UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Máster Universitario en Investigación e Innovación en Tecnologías de la Información y las Comunicaciones

–TRABAJO FIN DE MÁSTER–

Evaluación de un Sistema de Detección de Actividad de Voz

Autor: León Augusto Bourgeat Terán

Tesis de grado presentada como requisito para la obtención del Título de:
Máster Universitario en Investigación e Innovación en Tecnologías
de la Información y las Comunicaciones

(Master of Science)

Madrid, Septiembre 2013

Evaluación de un Sistema de Detección de Actividad de Voz

AUTOR: Augusto Bourgeat-Terán

TUTOR: Dr. Javier González-Domínguez

Máster Universitario en Investigación e Innovación en Tecnologías

de la Información y las Comunicaciones

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Madrid, Septiembre 2013

Departamento:	Tecnología Electrónica y de las Comunicaciones Escuela Politécnica Superior Universidad Autónoma de Madrid (UAM), ESPAÑA
Título:	Evaluación de un Sistema de Detección de Actividad de Voz
Autor:	**Augusto Bourgeat-Terán**
Director:	**Dr. Javier González-Domínguez** Universidad Autónoma de Madrid, ESPAÑA
Fecha:	Septiembre del 2013
Tribunal:	Presidente: **Dr. Joaquín González Rodríguez** Universidad Autónoma de Madrid, ESPAÑA Vocal 1: **Dr. Javier González-Domínguez** Universidad Autónoma de Madrid, ESPAÑA Vocal 2: **Dr. Doroteo Torre Toledano** Universidad Autónoma de Madrid, ESPAÑA Suplente: **Dr. Daniel Ramos Castro** Universidad Autónoma de Madrid, ESPAÑA
Calificación:	

The research described in this Thesis was carried out within the ATVS – Biometric Recognition Group at the Dept. of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid.

Dedicatoria

Con mucho cariño a mis padres que me dieron la vida y una carrera para mi futuro, a mi esposa Karina, que siempre ha estado apoyándome y brindándome todo su amor, a mis hijos Augusto, Cybelle y Thierry que son pedacitos de amor de mi corazón y a mis profesores por darme el tiempo para realizarme profesionalmente.

Agradecimiento

Deseo demostrar mi más sincero agradecimiento a todas aquellas personas que a continuación citaré y muchas de las cuales han sido un soporte muy fuerte en momentos de angustia y desesperación.

Primero, dar gracias a Dios, por estar conmigo en cada paso que doy, para fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía.

Agradezco, a Javier González-Domínguez, Ph.D., Tutor del Trabajo de Fin de Máster, a quien considero como un profesor excelente por lo que, para mí es un honor haber realizado este trabajo bajo su dirección, siempre guardaré mi gratitud y reconocimiento por haberme permitido estudiar en una Universidad tan prestigiosa como lo es la Universidad Autónoma de Madrid.

Agradezco, a los Docentes de la Universidad Autónoma de Madrid, por compartir sus valiosísimos conocimientos conmigo e inspirar en mi admiración. También a todos quienes conforman el Área de Tratamiento de Voz y Señales ATVS – Biometric Recognition Group.

A mi esposa, Karina, por darme su amor, y sobre todo, estabilidad emocional y sentimental, para poder culminar este objetivo.

A mis hijos Augusto, Cybelle y Thierry, por el apoyo y alegría que me brindan.

A mis padres y hermanas, a quienes agradezco de todo corazón por su amor, cariño y comprensión.

A Julio Arboleda, por la colaboración brindada.

Muchas gracias a todos.

Augusto Bourgeat

Este proyecto ha sido realizado en el Área de Tratamiento de Voz y Señales (ATVS) en la Escuela Politécnica Superior de la Universidad Autónoma de Madrid.

Resumen

En este Trabajo de Fin de Máster presento un Detector de Aprendizaje de Voz no Supervisado, que puede operar bajo el supuesto de que las primeras tramas de una expresión pueden ser voz/no-voz, en un ambiente de señal ruidosa sin supervisión a través de Modelos de Mezclas de Gaussianas, sustentado en la función de energía de las sub-bandas.

El algoritmo implementado se fundamenta en el método descrito por D. Ying et al., quien propone un VAD basado en un marco de aprendizaje no supervisado.

Algunas restricciones al modelo estadístico se introducen en el algoritmo, que da forma a las relaciones de distribuciones de voz y no-voz y mediante el uso de estas limitaciones, puede seleccionar uno o dos clases que se han modelado, de modo que el sistema no supervisado funcione correctamente. Las variables consideradas en el modelo son: longitud de las ventanas de inicialización, tamaño de cada ventana o frame, solapamiento, frecuencia de muestreo, longitud de la ventana FFT, números de canales Mel, calibración del umbral, umbral de votación, parámetros Hangover, coeficientes de pesos de actualización (alfa), y restricciones de medias (delta).

Se realizaron diversas pruebas con el fin de obtener resultados concretos y poder evaluar el rendimiento del VAD, bajo condiciones de voz o no-voz al inicio de las tramas; para lo cual se recurrió a las bases de datos MOBIO y TIMIT para evaluar el funcionamiento del algoritmo de Detección de Actividad de Voz, analizando la probabilidad de que pertenezca a una de las clases voz/no-voz para diferentes niveles de SNR.

Finalmente, se presentan las conclusiones extraídas a lo largo del trabajo junto con las propuestas del trabajo futuro.

Abstract

This work deals I present a Master's Thesis Voice Activity Detection Unsupervised, which can operate under the assumption that the first frame of an expression can be speech / non-speech in a noisy environment without supervision signal through Gaussian Mixture Model, based on the energy function of the sub-bands.

The algorithm implemented is based on the method described by D. Ying et al., who proposes a VAD based on an unsupervised learning framework.

Some restrictions statistical model are introduced in the algorithm, which gives shape to the distribution ratios voice and non-voice by using these limitations, may select one or two classes that are modeled so that the system unsupervised function properly. The variables considered in the model are: length initialization windows, each window size or frame, overlap, sampling frequency, FFT window length, Mel channel numbers, calibration of threshold, threshold voting, Hangover parameters, update weights coefficients (alpha), and restrictions of means (delta).

Various tests were performed in order to deliver concrete results and to evaluate the performance of the VAD, under conditions voice or non-voice at the beginning of the frames, for which was used to MOBIO and TIMIT databases to evaluate the performance algorithm for Voice Activity Detection, analyzing the probability that it belongs to a class speech / non-speech for different levels of SNR.

Finally, we present the conclusions of the work along together with proposals for future work.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

Glosario de Términos

ASR *Automatic Speech Recognition*(reconocimiento automático de habla).

ATVS *Área de Tratamiento de Voz y Señales (ATVS)* - Biometric Recognition Group de la Escuela Politécnica Superior de la Universidad Autónoma de Madrid, dedicada a la investigación en las áreas de biometría, reconocimiento de patrones, análisis de imágenes/voz y procesamiento de señales.

DFT *Discrete Fourier Transform* (transformada discreta de Fourier). Función de transformación ampliamente empleada en tratamiento de señales y campos afines para analizar las frecuencias presentes en una señal muestreada.

EER *Equal Error Rate* (tasa de error igual). Tasa de error, en los sistemas de reconocimiento biométrico, en que se igualan las tasas de falsa aceptación y falso rechazo.

FAR *False Acceptance Rate* (tasa de falsa aceptación). Porcentaje de errores, respecto al total de comparaciones realizadas, en los que se considera que el rasgo de test se corresponde con el patrón de referencia cuando en realidad corresponde a una identidad distinta.

EM *Expectation–maximization.* El algoritmo esperanza-maximización o algoritmo EM se usa en estadística para encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables.

FFT *Fast Fourier Transform* (transformada rápida de Fourier). Algoritmo para la implementación rápida de la DFT.

FRR *False Rejection Rate*(tasa de falso rechazo). Porcentaje de errores, respecto al total de comparaciones realizadas, en los que se considera que el rasgo de test se corresponde con el patrón de referencia cuando en realidad corresponde a una identidad distinta.

GMM *Gaussian Mixture Model* (modelo de mezcla de gaussianas). Técnica para el modelado de la identidad de un sujeto por medio del ajuste de un conjunto de gaussianas multivariadas a su distribución de características.

GSM *Global System for Mobile communications* El sistema global para las comunicaciones móviles.

HMM *Hidden Markov Model* (modelo oculto de Markov) Es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos de dicha cadena a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para llevar a cabo sucesivos análisis, por ejemplo en aplicaciones de reconocimiento de patrones.

IEEE *Institute of Electrical and Electronics Engineers*. Asociación técnico-profesional mundial dedicada a la estandarización. Promueve la creatividad, el desarrollo y la integración, comparte y aplica los avances en las tecnologías de la información, electrónica y ciencias para beneficio de la humanidad y de los profesionales.

ITU *International Telecommunication Union*. La Unión Internacional de Telecomunicaciones es el organismo especializado de Telecomunicaciones de la Organización de las Naciones Unidas encargado de regular las telecomunicaciones a nivel internacional entre las distintas administraciones y empresas operadoras

MAP *Maximum A Posteriori*. Técnica empleada para la adaptación de modelos de locutor a partir de un UBM en los sistemas basados en GMM.

MFCC *Mel Frequency Cepstral Coefficients* (coeficientes cepstrales en escala de frecuencias Mel). Coeficientes para la representación del habla basados en la percepción auditiva humana.

NIST *National Institute of Standards and Technology*(Instituto Nacional de Estándares y Tecnología de los Estados Unidos de América).

PDF *Probability Density Function* .Función de densidad de probabilidad, función de densidad, o, simplemente, densidad de una variable aleatoria continua describe la probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor.

Score Puntuación obtenida por un sistema de reconocimiento biométrico en la comparación entre un patrón de referencia y un rasgo biométrico de test.

SNR *Signal-to-Noise Ratio.* La relación señal/ruido se define como el margen que hay entre la potencia de la señal que se transmite y la potencia del ruido que la corrompe. Este margen es medido en decibelios.

SVM *Support Vector Machine* (máquina de vectores soporte). Clasificador discriminativo empleado en reconocimiento de locutor independiente de texto.

VAD *Voice Activity Detection.* La detección de actividad de voz, también conocida como detección de actividad de voz o de detección del habla , es una técnica utilizada en el procesamiento de voz en el que se detecta la presencia o ausencia del habla humana.

VoIP *Voice Over Internet Protocol.* Voz sobre IP es una metodología y un conjunto de tecnologías para la entrega de comunicaciones de voz y multimedia sobre protocolo de Internet (redes IP), como la Internet. Otros términos comúnmente asociados con VoIP son la telefonía IP , telefonía por Internet , voz sobre banda ancha (VoBB), la telefonía de banda ancha ,comunicaciones IP y servicios de telefonía de banda ancha.

Capítulo 1

Justificación y Objetivos

1.1. Introducción y Justificación

Una de las tareas fundamentales de aplicación de las tecnologías del habla es la detección de los períodos de voz y silencio en una señal de audio. Por ejemplo, en codificación un Detector de Actividad de Voz (VAD del inglés Voice Activity Detection), permite reducir la velocidad de trasmisión y realizar un mejor aprovechamiento del ancho de banda disponible. En etapas de reconocimiento automático de voz, locutor o idioma, el objetivo del detector será reducir la tasa de error del reconocedor degradado por la inclusión de fragmentos ruidosos.

Típicamente un VAD procesa la señal de entrada en segmentos cortos (frames) de tiempo de 10 a 40 ms y concomitantemente extrae características de los indicados frames. La decisión del VAD (etiquetas de voz o no voz del frame) se fundamenta en la información obtenida de estas características.

Los sistemas de reconocimiento automático de voz, locutor o idioma, se ven frecuentemente afectados por el ruido. Numerosas técnicas se han desarrollado para disminuir el efecto del ruido a diversas escalas, pre-procesado, caracterización o modelado, sobre la tasa de reconocimiento. El objetivo de éste trabajo es realizar la implementación de un detector de actividad de voz robusto al ruido. La tarea de clasificación de voz/no-voz no es tan trivial como inicialmente podría parecer y la mayor parte de los algoritmos para VAD fallan cuando el nivel de ruido de fondo se incrementa.[1]

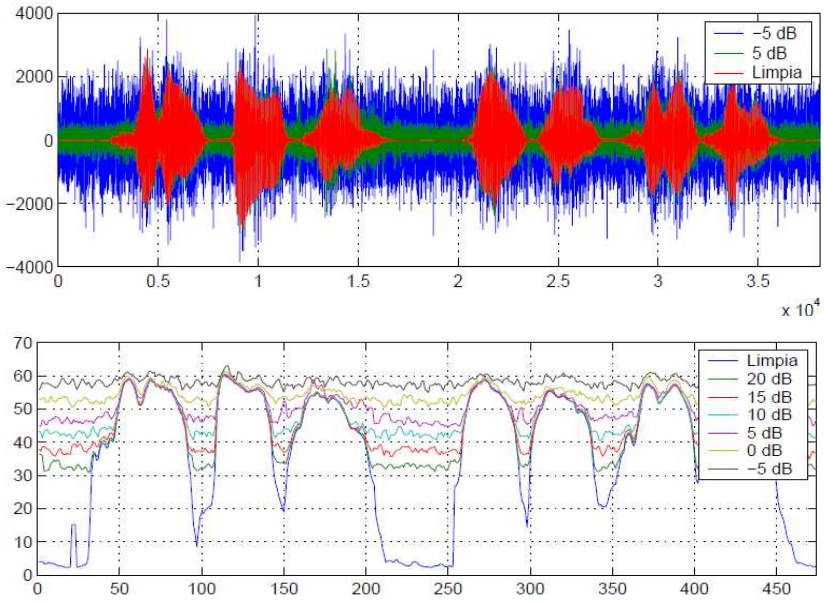


Figura 1.1. Niveles de energía para varias SNR. [1]

La efectividad del VAD es de gran importancia para la mayoría de las aplicaciones de las tecnologías del habla. En este trabajo se investigará las técnicas de Detección de Actividad de Voz y sus aplicaciones. Generalmente, los algoritmos, mejoran el comportamiento del detector en ciertos entornos de ruido, mientras que desmejoran en otros. La complejidad del tema hace

necesario, dedicar un esfuerzo importante en la mejora de la tecnología empleada en la detección de actividad y en estudios exhaustivo de nuevas técnicas de medida y evaluación de prestaciones.

1.2. Objetivos

El campo de la interacción oral hombre-máquina, en el que se incluyen los sistemas de reconocimiento automático del habla ASR (Automatic Speech Recognition), van ganando espacio como parte de los sistemas de nuevos productos comerciales.

Los sistemas ASR se caracterizan en modelos estadísticos de probabilidad entrenados previamente para representar las propiedades de la señal de voz, proporcionando resultados deseados en condiciones de laboratorio, pero en entorno reales, lo normal es que se ven afectados por el ruido. En escenarios reales la precisión de los sistemas de reconocimiento pierden efectividad, especialmente cuando disminuye la SNR.

El objetivo de los sistemas de VAD robustos al ruido es ser capaz de separar bien voz de no voz, con una alta tasa de acierto, inclusive cuando la calidad de la señal de voz de entrada se encuentre afectada por el entorno acústico. La investigación se centra en técnicas y algoritmos eficientes de los sistemas de reconocimientos.

Los objetivos específicos del trabajo serán:

- Estudiar y documentar el estado del arte sobre Detectores de Actividad de Voz y su utilización en aplicaciones de tecnología del habla.
- Formular e implementar un modelo de VAD para mejorar la robustez de la detección en entornos adversos.
- Analizar los resultados obtenidos en la implementación del modelo de estudio.

1.3. Motivación

El algoritmo implementado se fundamenta en el método descrito por D. Ying et al. [2], que se aplica para un VAD en un marco de aprendizaje no supervisado, algoritmo que se basa en Modelos de Mezclas Gaussianas Secuenciales (SGMM), y utiliza como parámetro la función de distribución de energía en las bandas de frecuencia Mel de la señal.

El VAD "Activity Detection Based on an Unsupervised Learning Framework" SGMM, se implemento considerando las conclusiones de los autores respecto al rendimiento del VAD y el nivel de análisis del algoritmo.

El análisis del algoritmo conlleva a una implementación práctica en un lenguaje de programación como por ejemplo C o MatLab.

Otra razón por la cual se decidió implementar el Modelo de Mezclas de Gaussianas Secuenciales, es porque está sustentado en modelos estadísticos que utiliza un umbral de decisión para la energía logarítmica de las sub-bandas, destacándose por su implementación en sistemas de detección y también por su aplicación en tiempo real.

Seleccionamos este modelo, por estar interesados en procesos no supervisados, en los cuales el VAD puede segmentar la señal sin importar que las tramas iniciales sean voz o no-voz, asimismo con cada una de las tramas que ingresan al VAD, permite la actualización del modelo, que se adapta a los umbrales de las condiciones de ruido.

He considerado utilizar el VAD en entorno ruidosos, empleando bases de datos de MOBIO y TIMIT y se ha replicado para aplicarlo a entornos reales y configurarlo para tal fin.

1.4. Detector de actividad implementado

Cómo construir modelos para discriminar voz/no-voz es un aspecto trascendental para los Detectores de Actividad de Voz (VAD). El aprendizaje semi-supervisado es la manera más común para la construcción de modelos de VAD.

En este trabajo se analiza la estructura de aprendizaje no supervisado en el diseño de un modelo estadístico para el VAD. Esta estructura está integrado por un Modelo de Mezclas de Gaussianas Secuenciales, que está conformado por un proceso de inicialización y un proceso de actualización.

En cada sub-banda, el Modelo de Mezclas de Gaussianas (GMM) se inicializa utilizando el algoritmo EM (Expectation–maximization) y sucesivamente se actualiza frame a frame. Cuando se utiliza el GMM, se calcula el umbral de autorregulación de discriminación para cada sub-banda. [2]

Con el objetivo de tener una visión global de la investigación, en la Figura 1.2. se presenta en forma esquemática las principales características del VAD.

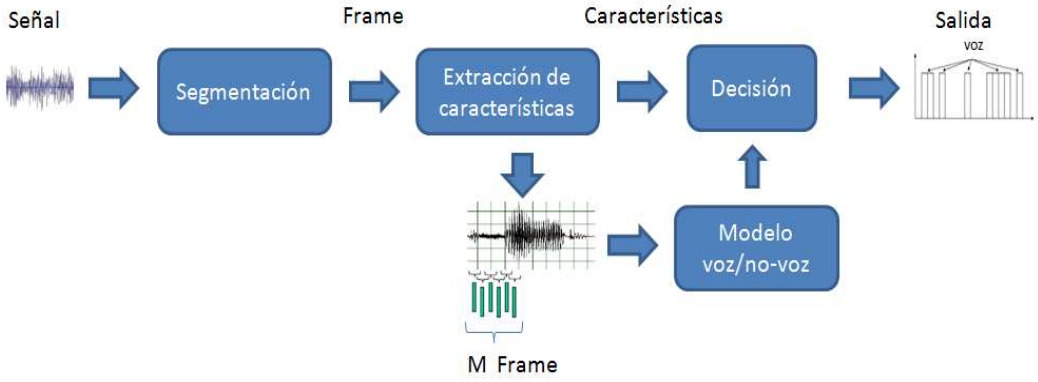


Figura 1.2. Esquema del método implementado.

El algoritmo analizado se basa en la estructura de modelado estadístico, que cuenta con tres propuestas para un VAD estadístico [3].

- El VAD analizado adopta el aprendizaje sin supervisión. Por lo cual, no se basa en el supuesto de "no habla al inicio de las tramas".
- Algunas restricciones a los modelos estadísticos se introducen en el modelo analizado. Se puede dar forma a las distribuciones de voz y no-voz; asimismo, mediante la aplicación de

estas limitaciones, el algoritmo analizado puede decidir automáticamente uno o dos grupos que se ha formado de modo que el sistema no supervisado funciona bien.

- La distribución de voz y no-voz serán aplicados en la decisión.

1.5. Estructura de la Tesis

En esta sección se presenta la estructura global de este trabajo describiendo el contenido de cada uno de los capítulos que se van a analizar:

- El Capítulo 2, presenta el Estado del Arte de la Detección de Actividad de Voz.
- El Capítulo 3, revisa los fundamentos del VAD basado en Modelos de Mezclas de Gaussianas Secuenciales.
- El Capítulo 4, analiza las bases de datos y el protocolo experimental.
- El Capítulo 5, evalúa el funcionamiento del algoritmo de Detección de Actividad de Voz, para diferentes niveles de SNR y experimentación en detección, utilizando las bases de datos MOBIO y TIMIT.
- En el último capítulo, presento las conclusiones, se resumen las aportaciones y finalmente se enumeran las posibles líneas futuras del trabajo de investigación.

Capítulo 2

Estado del Arte en Detección de Actividad de Voz

2.1. Introducción

Los Detectores de actividad de Voz se utilizan en numerosas áreas y aplicaciones de procesamiento de voz como: comunicación móvil [4][5], trasmisión de voz en tiempo real a través de Internet [6] y reducción de ruido en dispositivos de ayuda a la audición [5] [7]. El interés de los investigadores se ha centrado en el estudio de características robustas de la señal de voz y reglas de decisión basadas en modelos probabilísticos. [8]

2.2. Desafíos en el diseño del Detector de Actividad de Voz

Una primera aproximación de Detectores de Actividad de Voz, se basan en características de la señal como la energía, la periodicidad, las tasas de cruce por cero[9][10]; apoyadas en técnicas de discriminación de modelos heurísticos. Los VAD más sofisticados, utilizan estas características, pero basan su discriminación en modelos estadísticos[3]. Modelos estadísticos típicos están apoyados en clasificadores que asumen diversas distribuciones (ejemplo gaussianas) para describir las características de ruido y voz. Además de un buen rendimiento y consistencia a través del funcionamiento con varios tipos de ruido y SNRs, las características deseadas de un VAD incorporan baja complejidad computacional y adaptación rápida a los tipos de ruido cambiantes.

Ruidos fuertes y no estacionarios plantean desafíos a los sistemas de VAD, lo que ha obligado a que las investigaciones en los últimos años desarrollen sistemas robustos. Un sistema típico de VAD moderno utiliza fases de entrenamiento de mezclas de habla y ruido que son comparadas en la aplicación, requiriendo de etiquetados de la actividad de voz(aprendizaje supervisado)[11][12], o por lo menos en los datos de ruido similar a la del ruido encontrado en la aplicación (aprendizaje semi-supervisado). [13][14]

En este último caso, los métodos suponen que los datos de formación de ruido están disponibles porque requieren una inicialización de un modelo de ruido. Los métodos semi-supervisados se basan en supuestos paramétricos sobre el ruido (por ejemplo, Gaussianidad) que pueden ser alterados en ambientes de ruido no estacionario. [3]

Puede ser difícil obtener datos de entrenamiento especializados, por lo tanto, es deseable para diseñar un sistema de VAD sin supervisión, que funcione sin datos de entrenamiento y sea robusto, para que pueda ser utilizado en una variedad de entornos de ruido con amplios rango de señal-ruido. Los sistemas de VAD, como G.729B [16] [15] y AMR [17] , siguen un enfoque basado en reglas y por lo tanto no requiere datos de entrenamiento. Estos modelos han sido superados por métodos estadísticos y enfoques basados en clasificadores, que son más robustos y producen mejores resultados, pero requieren datos de entrenamiento etiquetados. Actualmente, se han interesado en desarrollar sistemas de VAD sin supervisión que tienen las ventajas de rendimiento de los sistemas supervisados. El enfoque habitual ha sido el de añadir un elemento de adaptabilidad a métodos supervisados y semi-supervisado existentes. [13]

Cuando se incrementa la potencia del ruido de fondo, un VAD debe afrontar los siguientes desafíos [5]:

Baja relación señal-ruido (SNR). La SNR es una medida genérica de como una señal ha sido empeorada por el ruido. SNR es definida como una relación de la potencia de la señal relacionado con la potencia de ruido [18]. Un VAD tiene que detectar voz o habla correctamente incluso si el ruido de fondo es muy fuerte o una persona está hablando en tono bajo. Este desafío es el más difícil de tratar en la práctica.

Variación rápidas del ruido de fondo. Adaptabilidad al ruido de fondo no estacionario, por ejemplo, en entornos de ruido de maquinaría pesada, arranque y parada aleatoria. [17]

Independencia del tipo de lenguaje, acento y voz. Un VAD debe tener el mismo rendimiento para contralto español (notas más graves de las mujeres) y el barítono italiano de los hombres. [5]

2.3. Componentes del Detector de Actividad de Voz

El diagrama de bloque de un detector de actividad está formado por los siguientes subsistemas. [9]

- **Señal de entrada**.- Conformada por la voz, el silencio y el ruido.

- **Pre-procesado.-** La señal de entrada pasa por un filtrado de reducción de ruido, mejorando la relación Señal/Ruido.

- **Extracción de características.-** Se obtiene las características de la señal (variaciones de energía, cruces por cero, correlaciones, etc.).

- **Clasificación.-** Se emplean métodos de clasificación de clases (voz/no-voz). Métodos sencillos a través de umbrales o métodos avanzados de clasificación como: Modelos Ocultos de Markov, Modelos de Mezcla de Gaussianas (GMM), Redes Neuronales Artificiales, Árboles de Decisión, etc.

- **Decisión.-** Se compara las características de la trama con los umbrales de los modelos, obteniéndose tramas de voz/no-voz (a nivel de trama) o a nivel de estructura tomando en cuenta conjuntos de resultados por trama para tomar decisiones a nivel de pulso.

- **Evaluación.-** Comparando los resultados obtenidos de la detección con los reales, procedentes de un etiquetado manual, se obtienen métricas: a) La precisión de marcas, b) Respuesta de las curvas ROC (Reciever Operating Characteristics) enfocado a pulsos de voz, es decir, las falsas aceptaciones y los falsos rechazos y c) Tasa de error de palabra.

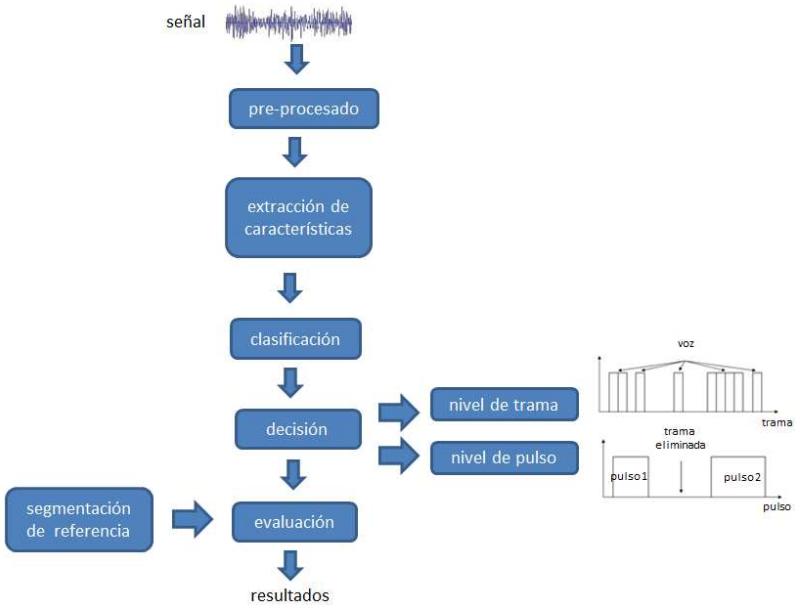


Figura. 2.1. Estructura del Sistema de Detección de Actividad de Voz.

2.3.1. Señal de voz

La entrada de un sistema de VAD es una señal de voz digitalizada con una frecuencia de muestreo. El IEEE define un canal de voz como un canal que es adecuado para la transmisión de voz o datos analógicos y tiene la máxima frecuencia utilizable de 300 a 3400 Hz. Aplicaciones de VAD se han utilizado ampliamente en sistemas de telefonía digital, donde la tasa de muestreo común es 8 KHz. Esto implica frecuencia digital máxima de 4 KHz. [18]

Segmentación:

La señal digital se procesa en cortas tramas de duración generalmente de 10-40 ms en aplicaciones de procesamiento del habla. Este período es tiempo suficiente para recopilar los datos necesarios para su procesamiento, pero es lo suficientemente corto como para que la señal de voz pueda considerarse estacionaria. Los frames están solapados con avance del frame igual al 50% o 33.3% de la duración de la trama.

2.3.2. Pre-procesado

En esta etapa se busca mejorar la relación voz/no-voz, con el fin de que el procedimiento de detección sea sencillo, fiable y robusto.

En entornos ruidosos, cuando la señal es pre-procesada, ayuda a que el detector de voz sea eficiente, pudiendo utilizar menos características en detectores robusto, o más fácil de determinar zonas de máximos y mínimos de energía, y en las zonas con mínimos de energía, se puede realizar una reestimación del ruido. [19]

Filtros Wiener

El estándar ETSI Aurora en pre-proceso para reducir el ruido, se basan en el análisis en frecuencia mediante filtros Wiener. Los filtros Wiener son usados para trabajar en ambientes ruidosos. [20]

Los filtros Wiener son filtros lineales de mínimos cuadrados que pueden ser usados para predicción, estimación, interpolación, filtrado de señal ruido, etc.

En tecnología digital, el diseño mediante filtros Weiner están presentes en aplicaciones de procesado de señal en receptores de comunicaciones, codificadores, etc.

2.3.3. Extracción de Características

Dado que los datos de entrada de audio no han sido procesados, son en gran medida redundantes y ruidosos para el procesamiento, por lo que las técnicas de extracción de características se utilizan para obtener la información esencial acerca de los datos que serían suficientes para su posterior procesamiento. El objetivo del extractor de características es el de obtener información comprimida de cada frame mediante el mapeo de sus datos a un vector de características. Las características llevan la información que debería ser suficiente para un VAD pueda clasificar la trama.

Algunas de las características que se han propuesto para un VAD incluyen tasa de cruce por cero [10], bandas de energía, envolvente de la señal, la entropía espectral, la divergencia espectral, coeficientes de frecuencia Mel cepstrum (MFCC) [21], etc.

2.3.4. Decisión

En esta etapa el VAD clasifica cada frame como voz o no voz (ruido). La regla de decisión podría ser simple (un umbral), así como muy complejo cuando se basa en clasificadores (máquina de vectores soporte (SVM), modelo oculto de Markov (HMM)) para producir la salida. [9]

La fase de decisión, comprende el proceso en el que se obtiene el resultado de clasificación. El resultado de esta clasificación puede realizarse en dos niveles: nivel de trama o de pulso.

La decisión a nivel de trama (Figura 2.2) implica la decisión de la trama actual sin tener en cuenta los resultados históricos de tramas anteriores o la estructura de lenguaje.

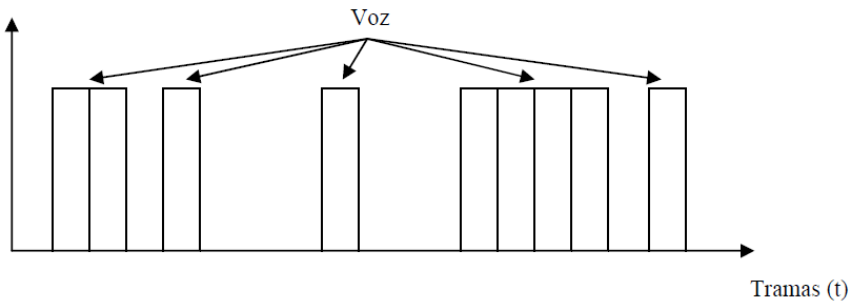


Figura 2.2. Decisión a nivel de trama.

La decisión a nivel de pulso toma en cuenta los resultados históricos de tramas anteriores y la estructura del lenguaje (Figura 2.3).

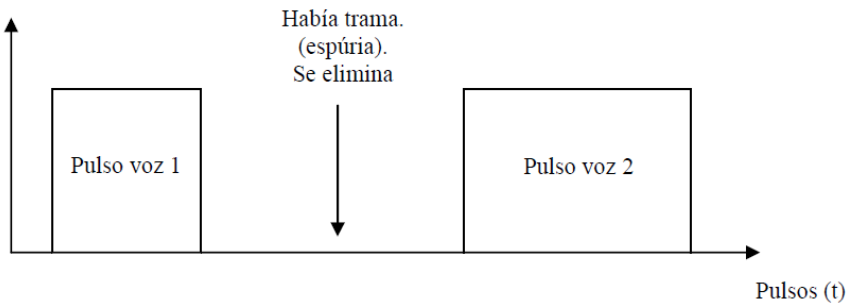


Figura 2.3. Decisión a nivel de pulso

En la Figura 2.3 se puede visualizar el significado de pulso de voz: agrupación de tramas que trata de simbolizar una pronunciación con un extremo inicial y otro final. De esta manera, la Figura 2.2 con decisión a nivel de trama se puede convertir en la Figura 2.3 a nivel de pulso; tramas con actividad aisladas suelen ser espurias (golpes, clicks), y tramas con actividad cercanas a varias tramas con actividad consecutivas pueden ser parte de una pronunciación.

Capítulo 3

Detector de Actividad de Voz no Supervisado

3.1.- Fundamentos del VAD basado en Modelos de Mezclas de Gaussianas Secuenciales

La función del detector de actividad de voz es diferenciar la voz activa de la no-voz en los audios. Desempeña un rol importante en los sistemas de comunicación como codificación de voz, reconocimiento de voz y mejora del habla. Su precisión incide en su rendimiento, especialmente en entornos adversos, un VAD robusto puede mejorar significativamente el rendimiento de los sistemas.[2].

Los VAD poseen características acústicas y modelos de discriminación. Los primeros algoritmos se relacionan con características acústicas robustas para distinguir la voz/no-voz. Las características de base energética son las más popular [22]. La relación de señal a ruido (SNR) se considera por lo general como una señal de energía para la discriminación [23].

La segunda característica popular es la cuasi-periodicidad de voz hablada [24]. Puede discriminar señal de voz del ruido de fondo.

Durante los últimos años, los VAD se han sustentado en modelos estadísticos para discriminar voz/no-voz; que en su mayoría se enfocan en la construcción de clasificadores de habla /no-habla. El clasificador clásico utiliza el modelo estadístico gaussiano para describir los coeficientes DFT [25], [26].

En estos modelos estadísticos, la razón de verosimilitud del discurso es la medida de clasificación, tienen una característica común, que generalmente se basan en el supuesto de que las expresiones siempre comienzan con señales de no habla.

Un modelo que inicia con señales de no habla se establece a partir de las primeras frames de un enunciado. Después de lo cual, los VAD discriminan cada frame que viene como voz/no-voz y luego retroalimenta el resultado de la discriminación para actualizar el modelo.

Las señales de no voz utilizada para la inicialización pueden ser generadas como muestras marcadas a mano, constituyéndose, esta inicialización del modelo en un proceso de aprendizaje supervisado; bajo el supuesto de la inicialización conocido como "principio de no habla".

El método de actualización basado en la retroalimentación se denomina toma de aprendizaje dirigido, que es una forma de realizar aprendizaje no supervisado.

Este enfoque de modelado que incorpora el aprendizaje supervisado con el no supervisado se denomina aprendizaje semi-supervisado. Constituyéndose en la característica común del modelo de construcción de los VAD.

Sin embargo, los VAD basados en aprendizaje semi-supervisado tienen un defecto en algunas aplicaciones prácticas. Si un enunciado inicia con una señal de voz, tal suposición no satisface, de manera que el modelo "principio de no habla" se inicializa con errores.

Un aprendizaje sin supervisión, puede operar bajo el supuesto de que las primeras frame de una expresión pueden ser voz/no-voz. La señal de voz ruidosa sin supervisión se agrupa en dos clases a través de algoritmos basados en la función de energía. Una clase con mayor promedio se considera voz, y la otra no-voz.

Las funciones de densidad de probabilidad (pdf) logarítmica de la energía del habla y no-habla son estimadas por el modelo basado en agrupación. Un umbral óptimo para la discriminación se encuentra en las pdfs. Los métodos de agrupamiento no supervisado tienen dos beneficios; en primer lugar, no es necesario para el diseño del VAD la suposición "principio de no habla", y en segundo lugar el umbral puede ser auto-regulado a los datos observados. [2].

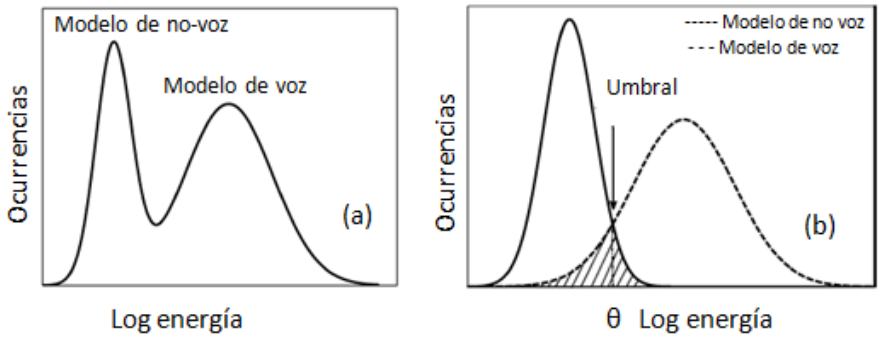


Figura. 3.1. Distribución de la energía logarítmica de una sub-banda de alta SNR. a) Distribución de habla ruidosa. b) Distribución de voz y no voz. [2]

Sin embargo, existe dos problemas que hay que resolver en estos VAD.

- Está ausente un mecanismo de actualización de los modelos, no son capaces de funcionar de manera online porque los algoritmos de clasificación se realizan normalmente de forma offline; por lo que no pueden aplicarse en sistemas de tiempo real.

- Es difícil para los VAD, decidir si uno o dos grupos han de ser formados. En caso de ausencia de habla o de baja SNR, las pérdidas de detección de voz es muy grave si se forman dos grupos.

Estas son las dos razones que hay que trabajar para el desarrollo de un VAD sin supervisión.

Considerando los problemas indicados, se implementa un VAD basado en una estructura de aprendizaje no supervisado. El GMM secuencial (SGMM) se utilizará para realizar este proceso de aprendizaje en cada sub-banda.

Las dos distribuciones del GMM, representan la probabilidad de voz y no-voz respectivamente. De acuerdo con el GMM, el umbral de autorregulación discrimina voz/no-voz en cada sub-banda. Los resultados de la discriminación de todas las bandas se resumen mediante el procedimiento de votación.

El algoritmo analizado se centra en la estructura de modelado estadístico, que cuenta con tres características que las diferencian de los VAD estadístico.

- El VAD analizado adopta el aprendizaje sin supervisión. Por lo tanto, no se basa en el supuesto de "no habla al principio de las tramas."

- Algunas restricciones a los modelos estadísticos se introducen en esta estructura. Por un lado, se puede dar forma a las relaciones de distribuciones de voz y no-voz; por otro lado, mediante el uso de estas limitaciones, el algoritmo puede decidir automáticamente uno o dos grupos que se ha formado de modo que el sistema no supervisado funcione bien.

- La distribución de voz y no-voz son considerados en la decisión.

3.2. Detalles de la estructura estadística para aprendizaje no supervisado del VAD

El VAD se define por características acústicas y clasificadores. Selecciona a la energía logarítmica de cada sub-banda como característica acústica. La señal de entrada se divide en varias sub-bandas Mel en el dominio de la frecuencia. La energía logarítmica se calcula utilizando el valor logarítmico de la suma magnitud en cada sub-banda. Con el tiempo, se suaviza para formar una envolvente para la clasificación.

Se construye dos modelos Gaussianos para describir las distribuciones de energía logarítmica de voz y no-voz. Estos dos modelos se incorporan en un GMM de dos componentes. Sus parámetros se estiman de forma no supervisada. La clasificación se lleva a cabo en cada sub-banda; y las decisiones de todas las sub-bandas se resumen en el procedimiento de votación.

En las siguientes sub-secciones, consideraremos un clasificador en una sub-banda.

3.3. Modelando la Distribución de Energía Logarítmica con GMM

Para el diseño del clasificador, iniciamos con la distribución de la energía logarítmica de una sub-banda. Considerando una sub-banda con una SNR alta, en la que la voz como no voz están presentes, la distribución de energía logarítmica en un período de tiempo dado se describe mediante un histograma, y se ilustra esquemáticamente en la Figura. 3.1 a).

Como la señal de ruido se supone que es más estacionaria que la señal de voz, la variación de la energía logarítmica de no-voz es más pequeña que la de la voz. Por lo tanto, existe un pico elevado correspondiente a un modo no voz, mientras que el pico plano representa el modo de voz.

"no-voz" indica la señal de ruido o silencio y "voz" expresa la superposición de señales de voz con ruido o solo voz. Observando que el promedio de energía de "voz" es mayor que el promedio de la energía "no voz", el pico de "no voz" está localizado en el lado izquierdo del pico de "voz". Esta combinación de distribución de probabilidades de voz y no voz, se denomina distribución bimodal.

Las energías de voz y no voz obedecen a distribuciones Gaussianas, la distribución bimodal puede ser construida por un GMM de dos componentes, identificando un componente con la media más pequeña como "no voz" y el otro componente para el modo de "voz".

Este modelo es descrito por las siguientes ecuaciones.

x_k : indica la energía logarítmica de una sub-banda en el tiempo k.

z : es la etiqueta voz/no-voz.

z pertenece {0 1}, donde "0" indica "no voz" y "1" indica "voz". De acuerdo con la regla de Bayes, tenemos la ecuación:

$$\mathrm{p}(\mathrm{x_k} \mid \lambda) = \sum_{z} \mathrm{p}(\mathrm{x_k}, \mathrm{z} \mid \lambda) = \sum_{z} \mathrm{p}(\mathrm{x_k} \mid \mathrm{z}, \lambda) \cdot \mathrm{p}(\mathrm{z}) \tag{1}$$

en donde:

$\mathrm{p}(z)$: es la probabilidad a priori de voz/no-voz, y en realidad es igual al coeficiente de peso ω_z . Además $\omega_0 + \omega_1 = 1$.

$\mathrm{p}(\mathrm{x_k} \mid \mathrm{z}, \lambda)$ representa la probabilidad de " x_k " dado el modelo de voz /no-voz.

$$\mathrm{p}(\mathrm{x_k} \mid \mathrm{z}, \lambda) = \frac{1}{\sqrt[2]{2\pi \cdot \mathrm{K}_z}} \exp\left\{-(\mathrm{x_k} - \mu_z)^2 / 2\mathrm{K}_z\right\} \tag{2}$$

μ_z y K_z representa la media y la varianza. $\lambda \equiv \{\mu_z, \mathrm{K}_z, \omega_z \mid z = 0,1\}$ son los parámetros del GMM.

$X \equiv \{x_1, x_2, x_3, x_4, x_5, \ldots x_M,\}$ es una secuencia de la energía logarítmica en una sub-banda. La pdf está dada por:

$$\mathrm{p}(\mathrm{x} \mid \lambda) = \prod_{k=0}^{M} p(\mathrm{x_k} \mid \lambda) \tag{3}$$

El parámetro λ se estima mediante la maximización de la función pdf. Desde el GMM, podemos obtener las pdfs de voz y no-voz de energía logarítmica, es decir $\mathrm{p}(\mathrm{x_k} \mid \mathrm{z}=1, \lambda) \cdot \mathrm{p}(\mathrm{z}=1)$ y $\mathrm{p}(\mathrm{x_k} \mid \mathrm{z}=0, \lambda) \cdot \mathrm{p}(\mathrm{z}=0)$. Estas dos pdf se muestran en la Figura. 3.1 b). De las dos pdf, se calcula un umbral óptimo θ para reducir al mínimo la clasificación de error. El umbral θ satisface:

$$\mathrm{p}(\theta \mid \mathrm{z}=1, \lambda) \cdot \mathrm{p}(\mathrm{z}=1) = \mathrm{p}(\theta \mid \mathrm{z}=0, \lambda) \cdot \mathrm{p}(\mathrm{z}=0) \tag{4}$$

La ecuación (4) es una ecuación de segundo grado con una incógnita θ. El umbral θ es una de sus raíces localizar entre las dos medias, a saber $\mu_{z=0} \le \theta \le \mu_{z=1}$. Las muestras con menor energía logarítmica que θ se clasifican como "no voz", y las muestras con mayor energía logarítmica que θ como "voz". La sombra en la Figura. 3.1 b) indica el error de clasificación.

3.4. Estimación del GMM Secuencial (SGMM)

El punto principal del modelo es estimar el conjunto de parámetros λ.

Mientras que el habla y el ruido son señal a trozos estacionarios, los parámetros de GMM deben ser adaptados a la variación de la señal.

La estimación consta de una inicialización fuera de línea y un proceso de actualización secuencial.

El GMM inicial se establece por el algoritmo EM, y posteriormente se actualiza incrementalmente con los próximos datos.

El conjunto de parámetros en el tiempo k se representa como:

$$\lambda_k \equiv \{\mu_{k,z}, \mathrm{K}_{k,z}, \omega_{k,z} \mid z = 0,1\}$$

λ_0 es el juego de parámetros inicial, estimado a partir de las primeras (M+1) muestras del algoritmo EM.

Las siguientes son las fórmulas de reestimación EM.

$$\overline{\omega}_{0,z} = \frac{1}{M+1} \sum_{j=0}^{\mathrm{M}} \mathrm{p}\left(z \mid \mathrm{x}_{\mathrm{j}}, \lambda'_0\right) \tag{5}$$

$$\overline{\mu}_{0,z} = \frac{\sum_{j=0}^{\mathrm{M}} \mathrm{x}_{\mathrm{j}} \cdot \mathrm{p}\left(z \mid \mathrm{x}_{\mathrm{j}}, \lambda'_0\right)}{(M+1) \cdot \overline{\omega}_{0,z}} \tag{6}$$

$$\overline{\kappa}_{0,z} = \frac{\sum_{j=0}^{\mathrm{M}} \left(\mathrm{x}_{\mathrm{j}} - \overline{\mu}_{0,z}\right)^2 \cdot \mathrm{p}\left(z \mid \mathrm{x}_{\mathrm{j}}, \lambda'_0\right)}{(M+1) \cdot \overline{\omega}_{0,z}} \tag{7}$$

donde:

$$\mathrm{p}\left(z \mid \mathrm{x}_{\mathrm{j}}, \lambda'_0\right) = \frac{\omega'_{0,z} \cdot \mathrm{p}\left(\mathrm{x}_{\mathrm{j}} \mid z, \lambda'_0\right)}{\sum_z \omega'_{0,z} \cdot \mathrm{p}\left(\mathrm{x}_{\mathrm{j}} \mid z, \lambda'_0\right)} \tag{8}$$

λ'_0 es el conjunto de parámetros antiguos y $\omega'_{0,z}$ son los coeficientes de ponderación de λ'_0.

$\overline{\lambda}_0 \approx \{\overline{\omega}_{0,z}, \overline{\mu}_{0,z}, \overline{\kappa}_{0,z}\}$ indica el nuevo conjunto de parámetros re establecidos a partir de λ'_0.

En la siguiente iteración, λ'_0 se sustituye por $\overline{\lambda}_0 \approx \{\overline{\omega}_{0,z}, \overline{\mu}_{0,z}, \overline{\kappa}_{0,z}\}$. Esta iteración continúa hasta que el algoritmo EM converge.

Al final $\overline{\lambda}_0$ es el parámetro inicial para valorar λ_0 que estamos resolviendo.

Dado λ_0, el umbral θ_0 se obtienen mediante el uso de la ecuación (4). Finalmente, las primeras (M+1) muestras se clasifican con θ_0.

Después de establecer el GMM inicial, el problema conlleva a la actualización de este modelo para los siguientes datos.

El esquema básico de la realización de un GMM secuencial es actualizar incrementalmente el conjunto de parámetros utilizando las últimas K muestras. Supongamos λ_K se conoce en el momento K+1, los parámetros en λ_{K+1} se derivan por las ecuaciones iterativas (9) - (12).

$$\omega_{k+1,z}=\frac{\sum_{j=k-K+1}^{k}\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_k\right)+\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{k+1}},\lambda_k\right)}{K+1} \tag{9}$$

$$\mu_{k+1,z}=\frac{\sum_{j=k-K+1}^{k}\mathrm{x}_{\mathrm{j}}\cdot\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_k\right)+\mathrm{x}_{\mathrm{k+1}}\cdot\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{k+1}},\lambda_k\right)}{\sum_{j=k-K+1}^{k}\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_k\right)+\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{k+1}},\lambda_k\right)} \tag{10}$$

$$\kappa_{k+1,z}=\frac{\sum_{j=k-K+1}^{k}\left(\mathrm{x}_{\mathrm{j}}-\mu_{k+1,z}\right)^2\cdot\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_k\right)+\left(\mathrm{x}_{\mathrm{k+1}}-\mu_{k+1,z}\right)^2\cdot\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{k+1}},\lambda_k\right)}{\sum_{j=k-K+1}^{k}\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_k\right)+\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{k+1}},\lambda_k\right)} \tag{11}$$

donde:

$$\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{k}},\lambda_k\right)=\frac{\omega_{k,z}\cdot\mathrm{p}\left(\mathrm{x}_{\mathrm{k}}\mid z,\lambda_k\right)}{\sum_{z}\omega_{k,z}\cdot\mathrm{p}\left(\mathrm{x}_{\mathrm{k}}\mid z,\lambda_k\right)} \tag{12}$$

El peso, la media, la varianza pueden ser considerados como los momentos de orden cero, uno y dos respectivamente de la energía logarítmica de voz/no-voz.

Este esquema no es tan deseable para el VAD. Por un lado, $\left\{\mathrm{p}\left(x_j\mid \mathrm{z}=1,\lambda_k\right)\mid j=k-K+1,\ldots,k\right\}$ se debe calcular en cada momento "k". El resultado es una pesada carga computacional. Por otro lado, no es beneficioso para el GMM el seguimiento de la variación de la señal debido a que tarde o temprano hacemos el muestreo con la misma contribución a la actualización de los modelos.

Basado en el esquema básico, se plantea el enfoque de la realización de un GMM secuencial. Supongamos que el GMM varía lentamente en el tiempo, y $\lambda_k\approx\lambda_{k+1}$ en el tiempo "k". Por consiguiente, $\sum_{j=k-K+1}^{k}\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_k\right)=\sum_{j=k-K+1}^{k}\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_{k-1}\right)$

La suma se aproxima al momento de orden cero, $\sum_{j=k-K+1}^{k}\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_{k-1}\right)\approx K\omega_{k,z}$.

De acuerdo con (9). Combinando estas relaciones, finalmente se obtiene la siguiente ecuación:

$$\sum_{j=k-K+1}^{k}\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{j}},\lambda_k\right)\approx K\omega_{k,z} \tag{13}$$

Sustituyendo (13) en (9), se obtiene:

$$\omega_{k+1,z}=\frac{K\cdot\omega_{k,z}+\mathrm{p}\left(z\mid \mathrm{x}_{\mathrm{k+1}},\lambda_k\right)}{K+1} \tag{14}$$

Si $\alpha = \frac{K}{K+1}$

$$\omega_{k+1,z} = \alpha \cdot \omega_{k,z} + (1-\alpha) \cdot p(z \mid x_{k+1}, \lambda_k) \quad (15)$$

Donde α puede ser considerado como un factor de olvido, donde 0 < α < 1; la condición de probabilidad $p(z \mid x_{k+1}, \lambda_k)$ se calcula mediante la ecuación (12).

La ecuación (6) puede ser aproximada por el momento de orden uno.

$$\sum_{j=k-K+1}^{k} p(z \mid x_j, \lambda_k) \cdot x_j \approx K\omega_{k,z}\mu_{k,z} \quad (16)$$

Reemplazando (16) en (10)

$$\mu_{k+1,z} = \frac{\alpha\omega_{k,z}\mu_{k,z} + (1-\alpha) \cdot p(z \mid x_{k+1}, \lambda_k) \cdot x_{k+1}}{\omega_{k+1,z}} \quad (17)$$

La ecuación (7) puede ser aproximada por el momento de orden dos

$$\sum_{j=k-K+1}^{k} p(z \mid x_j, \lambda_k) \cdot (x_j - \mu_{k+1,z})^2 \approx K\omega_{k,z}\kappa_{k,z} \quad (18)$$

Reemplazando (18) en (11)

$$\kappa_{k+1,z} = \frac{K\omega_{k,z}\kappa_{k,z} + (1-\alpha) \cdot p(z \mid x_{k+1}, \lambda_k) \cdot (x_{k+1} - \mu_{k+1,z})^2}{\omega_{k+1,z}} \quad (19)$$

El esquema secuencial consta de las ecuaciones (12), (15), (17), y (19), λ_{K+1} se deriva de λ_K y X_{K+1} en el tiempo k+1 es un proceso recursivo de primer orden. Luego, el umbral varia en el tiempo, θ_{K+1} se obtiene a partir de (4) dado λ_{K+1} .

Por último X_{k+1} es clasificado como voz/no-voz de acuerdo a $\theta_{K+1.}$

El GMM secuencial tiene dos ventajas sobre el básico.

- Como la probabilidad $\{p(x_j \mid z, \lambda_k) \mid j = k-K+1,...,k\}$ no es necesaria para cada tiempo "k", mejora mucho la eficiencia de cálculo.
- Los frames iniciales se olvidan con el transcurso del tiempo, y los frames actuales juegan un papel importante, por lo cual, la capacidad de seguimiento del algoritmo es más potente que el básico. (actualización selectiva)

El esquema presentado integra eficiencia computacional, rendimiento y requisitos de memoria.

3.4. Limitaciones para el GMM

En las sub-bandas de SNR alta, la distribución de energía logarítmica confirma la distribución bimodal.

Sin embargo, en las sub-bandas de SNR baja, la distribución de voz no es tan fácil de percibir en el histograma de energía logarítmica; especialmente cuando la señal de voz está ausente, sólo existe el histograma de no-voz. Esta distribución que comprende sólo un modo de no-habla se denomina distribución unimodal.

Si la distribución unimodal se construye por un GMM de dos componentes, ocasionarán errores de detección en valores perdidos. Por lo tanto, la propuesta de GMM de dos componentes debe mejorarse para armonizar correctamente la distribución unimodal con la bimodal.

Para tal fin, se establece un umbral δ utilizando la diferencia de medias; por consiguiente, una distribución bimodal debe satisfacer la relación:

$$\mu_{k,1} > \delta + \mu_{k,0}$$

Existe una relación similar entre la varianza de voz y no voz de la distribución bimodal. Bajo el supuesto de que la señal de ruido es más estacionaria que la señal de voz; las varianzas deben satisfacer la relación:

$$\mathrm{K}_{k,1} > \mathrm{K}_{k,0}$$

El conjunto de parámetros para la distribución bimodal puede ser estimado por el algoritmo EM, pero los parámetros del GMM para la distribución unimodal tienen que ser estimada de una manera especial.

En la distribución unimodal con SNR baja, todas las muestras pueden ser aproximadamente consideradas como no-voz. La media y la varianza de la componente no-voz pueden suponerse como la de todas las muestras.

Para la distribución unimodal, los parámetros de la componente de voz son imposibles de ser estimada a partir de los datos reales; por lo que debe construirse un componente virtual para la distribución de voz, donde su media y varianza son, respectivamente establecidas como $\delta + \mu_{k,0}$ y $\mathrm{K}_{k,0}$.

Tratándose de distribuciones bimodales o unimodal, los parámetros GMM deben satisfacer la relación

$$\mu_{k,1} = \max\left\{\mu_{k,1}, \delta + \mu_{k,0}\right\} \tag{20}$$

$$\mathrm{K}_{k,1} = \max\left\{\mathrm{K}_{k,0}, \mathrm{K}_{k,1}\right\} \tag{21}$$

donde δ cumple un compromiso entre componentes espectrales débiles del voz y componentes espectrales no-voz fuertes. En ambientes muy ruidosos, un elevado δ es beneficioso para

rechazar este último, mientras que una pequeña δ es ventajosa para detectar los primeros en entornos de bajo ruido.

Si el componente de voz virtual está construido por la distribución unimodal, el centro de distribución de la componente virtual $\delta+\mu_{k,0}$ se desvía lejos de todas las muestras. Como resultado, $p(x_k \mid z=0,\lambda_k) >> p(x_k \mid z=1,\lambda_k)$, y así, de acuerdo con (12), $p(z=0 \mid x_k,\lambda_k) >> p(z=1 \mid x_k,\lambda_k)$.

En el proceso de re-estimación, $\omega_{k,1}$ se aproximará a 0, por lo que los denominadores de (6) y (7) serán cero cuando se estima el componente de voz.

Por último, el algoritmo EM no converge si la restricción (20) se activa.

Este fenómeno también se produce en el proceso secuencial. Para protegerse contra ello, otra restricción debe ser considerada en los coeficientes de ponderación:

$$\omega_{k,1} = \max\{\omega_{k,1}, \varepsilon\}$$

$$\omega_{k,0} = 1 - \omega_{k,1} \qquad (22)$$

donde "ϵ" es pequeño y mayor que cero. En el algoritmo EM, la re estimación se terminará cuando esta limitación se activa.

Estas limitaciones provienen de las relaciones de distribución entre voz y no voz, que juegan diferentes papeles en esta estructura del VAD.

Además de la configuración de la relación de distribución, la restricción en las media (20) decide cuántos grupos se han de formar.

Cuando se activa, todas las muestras se agrupan en una clase, de lo contrario, dos clases se formarán. La restricción de pesos en (22) es un esclavo de ella. Seguirá la restricción de medias para ser activado. La función de restricción (23) da forma a la relación de la varianza. Todas las restricciones del GMM se incluyen en la inicialización y procesos de actualización.

Se diseñan dos pruebas, una en la sub-banda de SNR alta y otro en la sub-banda SNR baja, en un enunciado ruidoso para mostrar las funciones de las restricciones.[2]

La envolvente de una sub-banda de SNR alta se ilustra en la Figura. 3.2.

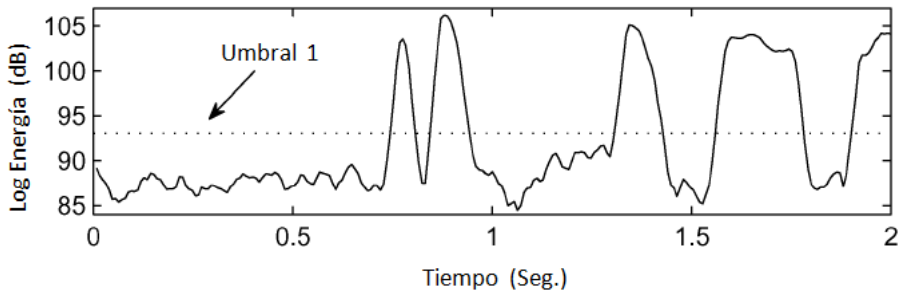


Figura. 3.2. Envolvente de la energía logarítmica de una sub-band de SNR alta.

La Figura. 3.3. modela su histograma, donde la energía logarítmica confirma la distribución bimodal.

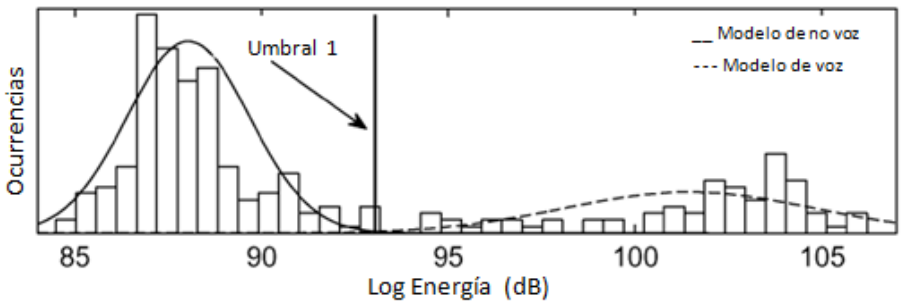


Figura. 3.3. Histograma de la energía logarítmica para una SNR alta.

A medida que la diferencia de medias es mayor que δ, las restricciones no se activan, y el umbral óptimo 1 se etiqueta en la Figura. 3.2.

Comparando el límite determinado con la correspondiente predicción en la Figura. 3.6, se puede ver que el clasificador funciona bien en la sub-banda con SNR alta, debiendo aclarar que no toda señal de voz está presente en esta sub-banda.

En una sub-banda con SNR baja como la de la Figura. 3.4.

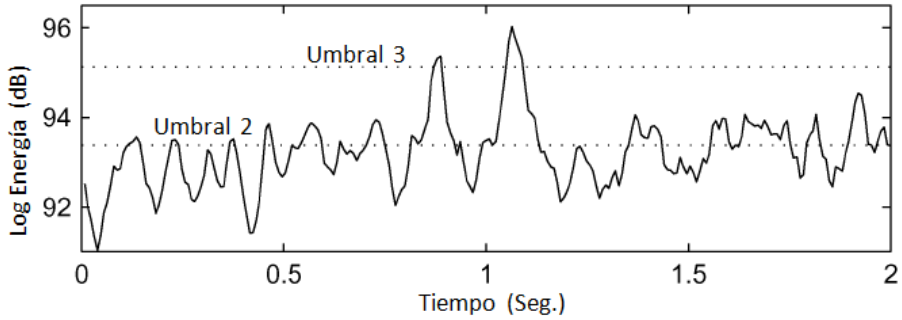


Figura. 3.4. Energía logarítmica de una sub-banda con SNR baja.

El histograma en la Figura. 3.5. (a) y (b) ilustra que la energía logarítmica sigue una distribución unimodal. En primer lugar, se ajusta a esta distribución sin restricciones, como se muestra en la Figura. 3.5. (a), y luego se muestran con restricciones, Figura. 3.5. (b).

Los límites encontrados se etiquetan en la Figura. 3.4, y se ilustran en la Figura. 3.6, donde el límite de 2 y 3 provienen de la banda de SNR baja y el límite 1 de la banda de SNR alta.

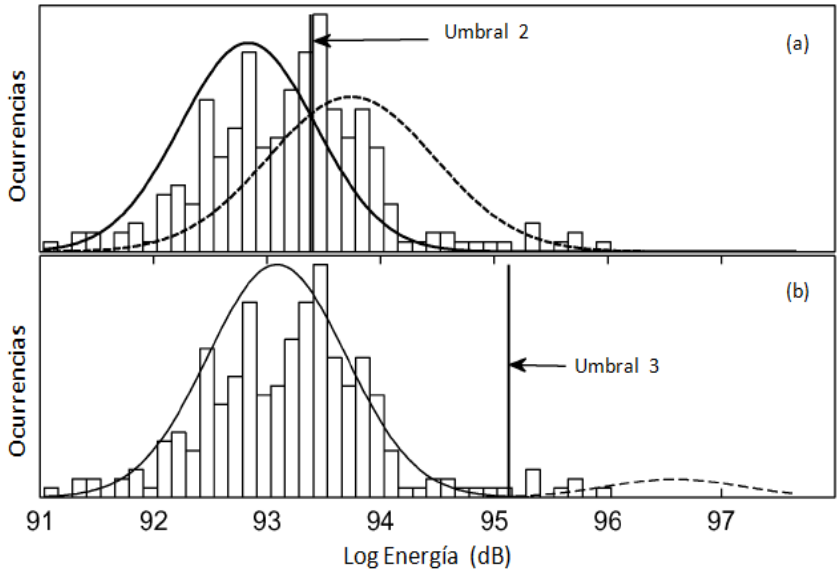


Figura. 3.5. Histograma de la energía logarítmica para una SNR baja.

Se puede observar que los valores perdidos de voz es muy grave sin restricciones, mientras que es menos probable que ocurra con limitaciones. A partir de estas pruebas, se puede constatar que el algoritmo EM con restricciones es aplicable tanto con la distribución bimodal como con la unimodal.

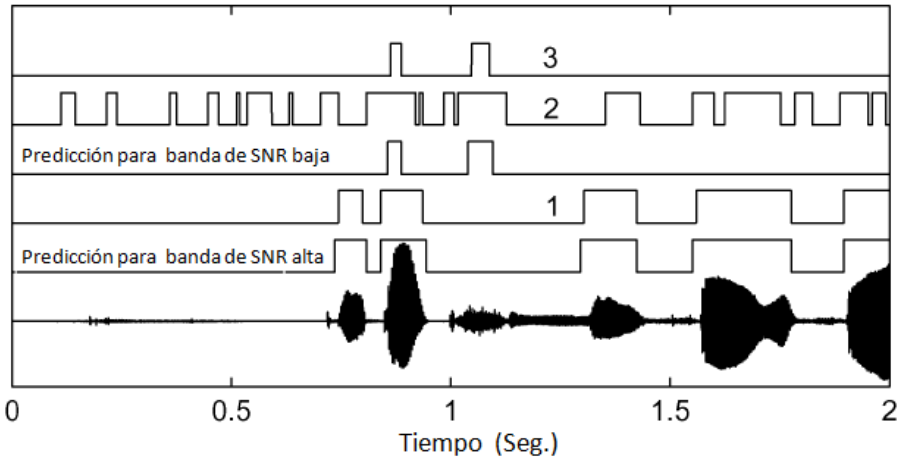


Figura. 3.6. Límites de VAD determinado por los umbrales 1, 2, y 3.

3.6. Implementación del sistema del VAD

El análisis muestra la clasificación de voz/no-voz en una sola sub-banda. La decisión basada en una sub-banda no puede producir un VAD fiable, por lo que deben combinarse todas las decisiones de las sub-bandas en un procedimiento de votación.

Vale la pena mencionar la razón para el tratamiento de cada sub-banda individualmente con un GMM univariado en vez de forma conjunta con un GMM multivariable; siendo la razón principal

el aprendizaje no supervisado, donde los modelos agrupados se deben identificar como voz/no-voz de acuerdo a sus características estadísticas.

Con el GMM univariado, la identificación del clúster se puede hacer fácilmente mediante la comparación del valor escalar de las dos medias del clúster. Si cada frame se toman como una unidad de decisión mediante el uso de los GMM multivariado, el procedimiento de votación sería innecesario; sin embargo, no hay garantía de que todos los elementos de las medias del clúster son mayores que las de otra. Como resultado, los grupos son difíciles de ser identificados basándose en el vector media. Por esta razón, el enfoque de " GMM univariado + procedimiento de votación " es más apropiado que el GMM multivariante.

3.7 Descripción del VAD basado en Modelo de Mezclas de Gaussianas Secuenciales

El VAD basado en el Modelo de Mezclas de Gaussianas Secuenciales funciona en tiempo real, consta de los siguientes pasos:

1. Para las Primeras (M+1) frames.

2. Para cada Sub-banda Mel.

3. Extraer la envolvente del logaritmo de la energía.

4. Establecer un GMM utilizando EM con restricciones.

5. Determinar el umbral del GMM aplicando la ecuación 4.

6. Calibrar el umbral mediante la ecuación 24.

7. Clasificar las (M+1) muestras como voz/no-voz.

8. Fin

9. Resumir todas las sub-bandas, seleccionándolas por votación.

10 Clasificar voz/no-voz utilizando el esquema hangover.

11 Fin

12 Para cada nueva frame en el tiempo (k+1).

13 Encontrar FFT y calcular x_{k+1} de cada sub-banda Mel.

14 Para x_{k+1} de cada sub-banda.

15 Calcular $p(z|x_{k+1}, \lambda_k)$ aplicando la ecuación 12.

16 Actualizar los coeficientes de pesos utilizando la ecuación 15.

17 Analizar las restricciones de los pesos mediante la ecuación 22.

18 Actualizar la media con la ecuación 17.

19 Analizar las restricciones de las medias mediante la ecuación 20.

20 Actualizar la varianza aplicando la ecuación 19.

21 Analizar las restricciones de las varianzas mediante la ecuación 21.

22 Determinar el umbral del GMM utilizando la ecuación 4.

23 Calibrar el umbral utilizando la ecuación 24.

24 Clasificar x_{k+1} como voz/no-voz.

25 Fin

26 Resumir todas las sub-bandas, seleccionándolas por votación.

27 Discriminar el (k+1) frame por el esquema hangover.

28 Fin

3.7.1. Energía logarítmica de una sub-banda

Se inicia extrayendo la envolvente de la energía logarítmica de una sub-banda; la señal de voz con ruido se divide en tramas o frame mediante la utilización de una ventana de Hamming. A continuación, se transforma al dominio de la frecuencia mediante la transformada rápida de Fourier (FFT), agrupándose en N sub-bandas de escala Mel.

Para la sub-banda, se calcula la energía logarítmica, aplicando la ecuación:

$$\bar{x}_k = 10 \cdot \log_{10}\left[\frac{1}{(f_{l+1} - f_l)} \sum_{j=f_l}^{f_{l+1}-1} \left|Y_{k,j}\right|^2\right] \tag{23}$$

dónde $Y_{k,j}$ es el **j** ésimo coeficiente de la DFT de la **k** ésima frame, y f_l es la frecuencia de índice correspondiente a la l ésima sub-banda de la escala Mel, l= 0,1,2,3,4,5, ...N.

El VAD utiliza características estáticas (logaritmo de la energía), y es entrenable de tal manera que el modelo se pueden adaptar a los diferentes canales de telefonía, por ejemplo de telefonía fija, GSM o voz IP.

3.7.2. El umbral de la sub-banda de la Energía logarítmica

El umbral de la sub-banda se calibra, para reducir al mínimo el error de clasificación, en favor de la perdida de voz, el sistemas VAD debe enfatizar más en la precisión de la detección de la Voz o habla. Para un clasificador binario, es conocido que la precisión de la detección de una clase se puede mejorar sacrificando la precisión de la otra clase, reduciéndose el umbral para mejorar la precisión de la voz, sacrificando la precisión de no-voz:

$$\theta_k' = \gamma \cdot (\theta_k - \mu_{k,0}) + \mu_{k,0} \tag{24}$$

dónde $0 < \gamma \leq 1$.

3.7.3. Post-Procesado de la Decisión (Suavizado)

La decisión binaria se realiza frame a frame. Esta decisión basada en segmentos no tiene en cuenta la naturaleza continua de la "Voz" y del "silencio". Por lo tanto, cualquier suavizado en la decisión puede aplicarse para prevenir recortes de voz o segmentos de ruido que posean características similares a la voz (picos), que son en realidad producidos por un ruido impulsivo. Una versión sencilla de suavizado para eliminar recortes consiste en extender la decisión en la detección de voz a los frame futuros, lo que produce un aumento en la tasa de acierto de voz y una disminución en la tasa de silencio.

3.7.3.1 Sistema Hangover

El sistema Hangover se utiliza para evitar pérdidas de discurso, mediante la reducción del riesgo de una parte de baja energía de la señal de voz falsamente rechazada.

En una implementación práctica, que requiere de un esquema de supresión para reducir la probabilidad de falsos rechazos [27]. El esquema Hangover considera la reducción del riesgo de una parte de baja energía de la palabra al final de una expresión falsamente rechazada. Esto se basa en la idea de que los sucesos del habla están altamente correlacionados con el tiempo. Un esquema de rechazar se puede implementar como una máquina de estados, para su correcta visualización.

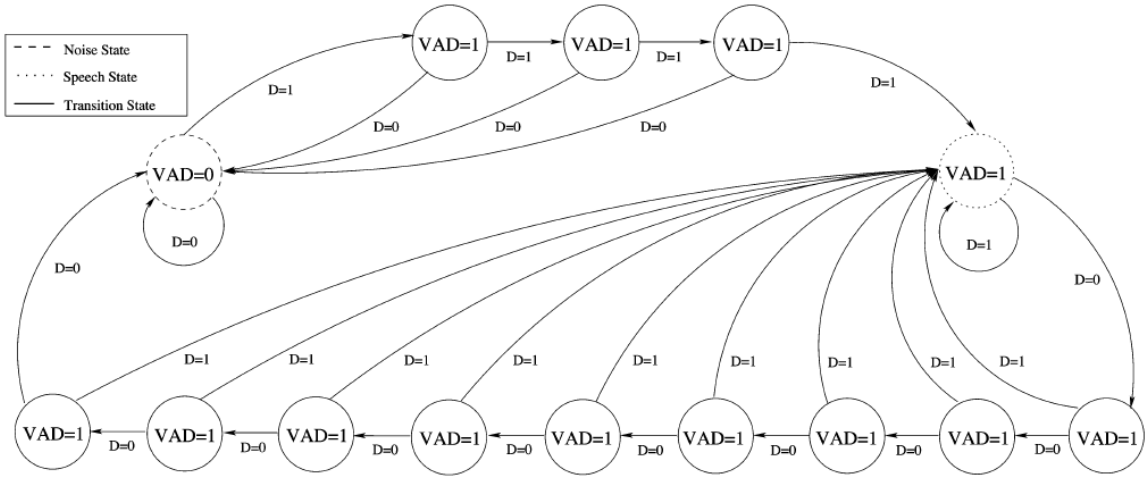


Figura 3.7. Máquina de estados esquema Hangover [27].

Capítulo 4

Bases de datos y protocolo experimental

4.1 Introducción

Para evaluar el funcionamiento del algoritmo de Detección de Actividad de Voz, se analiza la probabilidad de que pertenezca a una de las clases voz/no-voz para diferentes niveles de SNR, para lo cual se recurre a las bases de datos MOBIO y TIMIT. A estas bases de datos de señales de audio que no contienen ruido se les ha añadido tres ruidos reales mezclados artificialmente a la voz con SNR de 10, 5 y 0 dB, que han sido grabados en escenarios comunes para terminales de telecomunicaciones.

Un protocolo experimental se define como el conjunto de condiciones que se aplican, así como también la base de datos a utilizar en el análisis del VAD.

4.2. Bases de datos

Se utiliza la base de datos MOBIO y TIMIT para los pruebas; la base de MOBIO son ficheros con entornos de conversación de fondo y TIMIT son ficheros de voces limpias. El ruido se añade digitalmente terminando con SNR de 10 dB, 5dB y 0dB.

4.2.1. Base MOBIO

MOBIO (MObile BIOmetric consortium) es una gran base de datos de teléfono celular bi-modal (audio/visual), que incluye datos de discurso y datos faciales divididos en sesiones. A partir de esta base de datos, se extrae la señal de voz. [28].

La base de datos se registró de la siguiente manera:

- Set de respuestas: a los usuarios se les hicieron preguntas: ¿Cuál es su nombre? (alrededor de 7 segundos).
- Lectura de un discurso: los usuarios leen 3 oraciones fijas (máximo 30 segundos).
- Libertad de expresión: los usuarios proporcionan cinco hasta diez segundos, respuestas a las preguntas al azar.

Los ficheros de voz se nombran en función de las características de los locutores. Al Detector de Actividad de Voz se pasan los ficheros de audio con ruido y con SNR de 10dB, 5dB y 0dB.

4.2.2. Base TIMIT

La base TIMIT es un conjunto de datos de voz para estudios fonético-acústicos, desarrollo y evaluación de sistemas de reconocimiento automático del habla [29] y cada elemento transcrito se ha delineado en el tiempo. TIMIT es un corpus de fonético y léxico de discurso transcrito del inglés americano de diferente genero y dialectos. TIMIT fue diseñado para profundizar en el conocimiento fonético-acústico y sistemas de reconocimiento automático del habla, encargado

por DARPA a Texas Instruments (TI) y al Instituto Tecnológico de Massachusetts (MIT), de ahí el nombre del corpus. Un ejemplo de un discurso TIMIT:

```
3050      5723    she
5723      10337   had
9190      11517   your
11517     16334   dark
16334     21199   suit
21199     22560   in
22560     28064   greasy
28064     33360   wash
33754     37556   water
37556     40313   all
40313     44586   year
```

4.3. Etiquetado manual para base MOBIO

Una vez que se han seleccionado los archivos de audio de la base MOBIO, que contienen frases, palabras y expresiones del lenguaje, se procede a la identificación manual de voz y no voz; proceso conocido como etiquetado.

El etiquetado diferencia en un archivo de sonido las frases de voz, marcando la etiqueta "1" y para el caso de silencio o ruido la etiqueta "0"; de esta manera se puede diferenciar entre voz y no voz en al archivo de audio.

El etiquetado son marcas en el tiempo que nos indican donde empieza y termina una frase completa de habla. En el proceso de etiquetado se alinea en el tiempo la duración de la voz y del silencio presentes en los archivos de sonido.

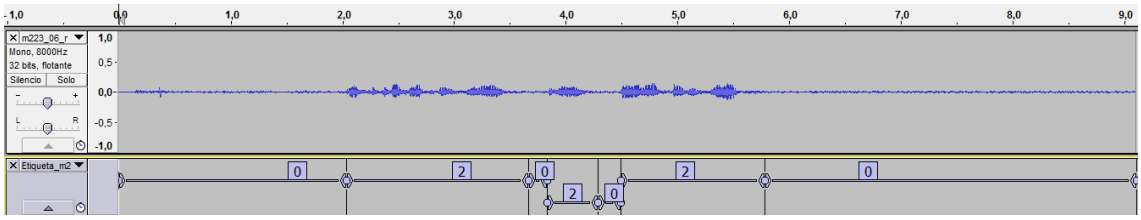


Figura 4.1. Etiquetados manual de la señal de audio m223_06_r10_i0_0.wav.

El conjunto de datos utilizado de la base MOBIO, consta de 44 archivos de audio, que contienen frases, palabras y expresiones del lenguaje, proceso de identificación manual de voz y no voz a través de etiquetas verdaderas realizadas manualmente; se generan archivos de audio de prueba, con SNR de 10dB, 5dB, 0dB, en tres escenarios diferentes de voz contaminados con ruido de conversaciones de fondo.

4.4. Etiquetado para la base TIMIT

Si elaboran las etiquetas reales de los archivos de audio para la base TIMIT. Un ejemplo de la formación de estas etiquetas es el siguiente:

Por ejemplo el archivo de audio: test_dr1_mwbt0_sx23.wav, se tiene el archivo:
test_dr1_mwbt0_sx23.wrd, en el siguiente formato:

Tabla 4.1. Archivo de etiquetado TIMIT

Muestra Inicial	Muestra Final	Expresión
2353	7101	Those
7800	21040	Musicians
21040	31750	harmonize
32583	44920	marvelously

La columna 1 indica en que muestra empieza la expresión, la columna dos indica en que muestra termina la expresión; para encontrar los tiempos dividimos para 16000 que es la frecuencia de muestreo y tenemos en segundos los tiempos de cada expresión:

Tabla 4.2. Tiempos de etiquetado TIMIT

Tiempo inicial	Tiempo final	Expresión
0.147063	0.443813	Those
0.487500	1.315000	musicians
1.315000	1.984375	harmonize
2.036438	2.807500	marvelously

Se construye las etiquetas teniendo cuenta el número de muestras del archivo de audio (59597):

```
0.0000000   0.1470625   0
0.1470625   0.4438125   1
0.4438125   0.4875000   0
0.4875000   1.9843750   1
1.9843750   2.0364375   0
2.0364375   2.8075000   1
2.8075000   3.1680000   0
```

Donde "0" indica no-Voz y "1" indica voz.

Se inspeccionó visual y audible para confirmar el verdadero etiquetado con el fin de lograr un buen rendimiento en los datos dados.

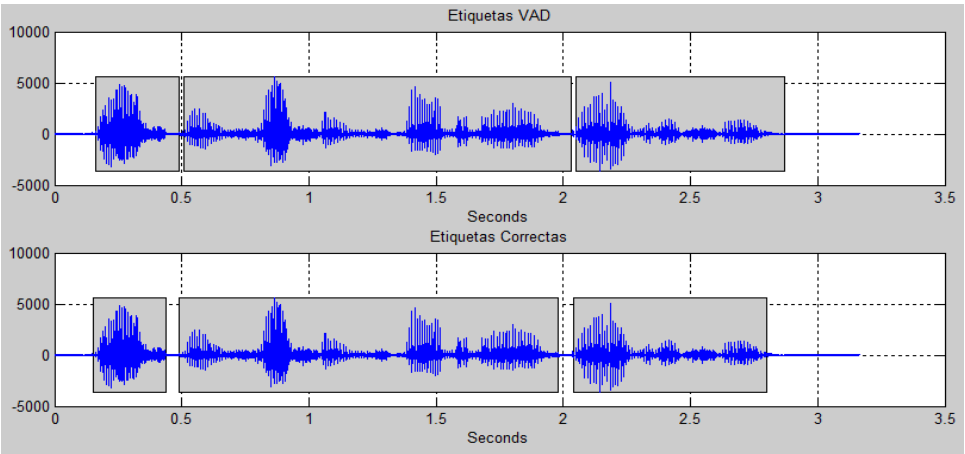


Figura 4.2. Etiquetas VAD y Etiquetas Correctas, inicio con tramas no-voz y SNR alta.

El conjunto de datos utilizado de la base TIMIT, consta de 23 archivos de audio, con una duración aproximada de 322 segundos; se generan archivos de audio de prueba, con SNR de 10dB, 5dB y 0dB, en tres escenarios diferentes de voz perturbados con ruido de conversaciones de fondo.

4.5. Añadir ruido a la señal de audio

Para los experimentos se ha añadido ruido a los archivos de audio de prueba de las bases MOBIO y TIMIT. Los ruidos se toman de muestras de la base de datos en formato RAW, muestreadas a 8 KHz. El ruido se añade a la señal de voz mediante los siguientes pasos:

- Un segmento de ruido se añade de acuerdo con la longitud de la señal de audio.
- La amplitud del segmento de ruido se regula según la relación SNR deseada.
- La señal de ruido se añade a la señal limpia para conseguir la voz distorsionada.

AddNoisPSO es una función implementada en MatLab, que añade ruido a un archivo de audio para una determinada relación SNR. Los niveles de voz y no-voz se determinan ponderando sofométricamente (UIT-T Recom. O.41). El nivel de voz se mide de acuerdo con la Recomendaciones UIT-T. P.56.

AddNoisePSO(cleanfile, noisefile, noisyfile, snrdB, seed)

donde:

- **cleanfile:** nombre del archivo de voz (RAW, 16-bit, 8 kHz)
- **noisefile:** nombre del archivo de ruido (RAW, 16-bit, 8 kHz)
- **noisyfile:** es el nombre del archivo de salida de audio con ruido (RAW, 16-bit, 8 kHz)
- **snrdB:** es el nivel de SNR (dB) deseado
- **seed:** es la semilla del generador aleatorio de reproducibilidad.

En nuestro experimento, hemos elegido tres muestras de ruidos: noiseA01.raw, noiseA07.raw y noiseB11.raw, que han sido elegido en forma aleatoria. Los ruidos corresponden a escenarios de fondo de personas que conversan en salones, en la calle, etc.

Los experimentos se llevan a cabo con cuatro diferentes niveles de ruido: Señal sin añadir ruido, alto (10 dB), medio (5 dB) y bajo (0 dB). Con estas señales de audio se medirá el rendimiento del VAD y la precisión en la detección de voz y no voz al comparando con las etiquetas verdaderas realizadas manualmente.

Tabla 4.3. Escenarios de audios experimental

Escenarios	**Relación Señal Ruido**		
	Señal sin aplicar ningún SNR		
Ruido 1	alto (10 dB)	mediano (5 dB)	bajo (10 dB)
Ruido 2	alto (10 dB)	mediano (5 dB)	bajo (10 dB)
Ruido 3	alto (10 dB)	mediano (5 dB)	bajo (10 dB)

4.6. Medidas de Evaluación

Se va a definir cuáles son las medidas de evaluación y de qué forma se calculan las métricas y errores que se presenta más adelante.

4.6.1. Tasa de acierto de voz y tasa de acierto de no-voz

El rendimiento VAD se mide por su precisión en la detección de voz como voz, y en la detección de ruido y pausas como no-voz. Estas medidas se definen como la tasa de aciertos de voz y tasa de aciertos de no de voz respectivamente, las cuales caracterizan los diferentes errores y evalúan el rendimiento del VAD.

La medida más significativa en esta distribución es la media de tasa de aciertos voz/no-voz. Esta medida se calcula mediante un promedio simple de las tasas de éxito de la voz y no-voz para un tipo de ruido en una determinada SNR.

4.6.2 Tasas de valores perdidos de voz y Tasas de falsas alarmas

Consideremos un VAD como una caja negra, que imprime etiquetas binarias, que indican segmentos de voz y no-voz de una señal de audio de entrada. La verdaderas etiquetas también se compone de ceros y unos. Para medir el rendimiento del VAD se comparan sus resultados con la realidad; siendo una forma de presentar la clasificación prevista y real de un VAD, la matriz de confusión [31].

Las relaciones entre los valores de la matriz de confusión son [5]:

Tabla 4.4. Matriz de Confusión del VAD.

		Etiquetas Reales	
		Voz (1)	No-Voz (0)
Salida del VAD (Predicted Labels)	Voz (1)	Verdaderos Positivos	Falsos Positivos
	No-Voz (0)	Falsos Negativos	Verdaderos Negativos

- La suma de todos los valores de la matriz de confusión, corresponde al número total de tramas etiquetadas.
- # de verdaderos positivos + # falsos negativos = # de frames etiquetadas como voz verdadera.
- # de falsos positivos + # verdaderos negativos = # de frames etiquetadas como no voz verdadera.
- # de verdaderos positivos + # de falsos positivos = # de frames etiquetados como voz del VAD.
- # de falsos negativos + # de verdaderos negativos = # de frames etiquetados como no voz por el VAD.

$$MR = \frac{\text{Falsos Negativos}}{\text{Falsos Negativos} + \text{VerdaderosPositivos}} \quad (25)$$

La tasa de falsas alarmas (FAR), muestra la proporción de datos no-voz clasificados como voz.

$$FAR = \frac{\text{Falsos Positivos}}{\text{Falsos Positivos} + \text{VerdaderosNegativos}} \quad (26)$$

La tasa de valores perdidos (RM) muestra la cantidad de datos de voz perdidas por un VAD. Una baja MR es crucial para aplicaciones que requieren un análisis total de los datos.

La tasa de valores perdidos se pueden calcular como (100% - tasa acierto voz) y tasa de falsas alarmas está definida como (100% - tasa acierto de no voz).

4.7. VAD de referencia a comparar

Para contrastar los resultados obtenidos con el VAD_{SGMM}, se realizan comparaciones con el VAD basado en Energía (VAD_Energía).

En el grupo ATVS se tienen un Sistema de Detección de Actividad de Voz, basadas en la fusión de dos VADs, para discriminar voz o silencio [32]. El primer VAD_1, trabaja por umbral de energía fijo, la energía de la señal filtrada (filtro Wiener) es comparada con un umbral; se tiene voz si supera el umbral. El segundo es el Detector de Actividad de Voz del SoX (Sound eXchange) VAD_SoX [33], utiliza la potencia Cepstral de una forma sencilla para detectar la voz. El algoritmo (internamente) emplea estimación adaptativa de ruido.

El VAD basado en Energía aplica la regla de decisión binaria AND, que involucra a los dos VAD y selecciona una alternativa, en el caso que los dos coinciden con su decisión de voz.

Tramas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
VAD_SoX	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1
VAD_1	1	1	0	0	0	0	0	0	1	1	0	1	1	1	1	1	0	0	0	0
VAD_Energía	1	1	0	0	0	0	0	0	1	1	0	1	1	1	1	1	0	0	0	0

Capítulo 5

Experimentos del sistema

5.1. Introducción

En este capítulo se obtienen resultados experimentales para el VAD basado en SGMM usando las bases de datos MOBIO y TIMIT, métricas utilizadas.

El VAD analizado utiliza un umbral ajustable para clasificar las tramas de voz y de ruido, se trata de un VAD de configuración robusto al ruido y realizable en tiempo real.

La Figura 5.1. a) muestra un fichero de audio de la base MOBIO con voz nítida (25dB) y la Figura 5.1. b) muestra un audio mezclado con ruido de fondo estacionario (0dB).

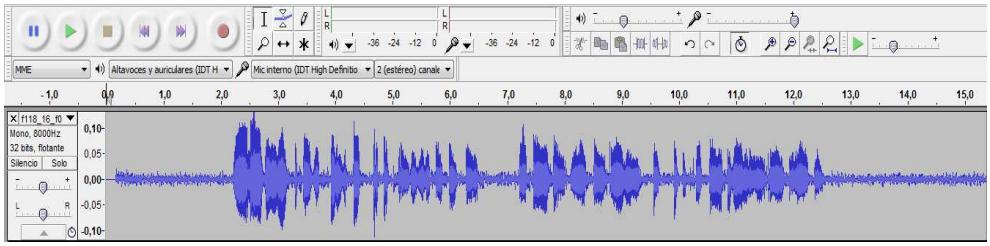


Figura 5.1. a) Fichero de voz limpia analizado (SNR= 25 dB), MOBIO.

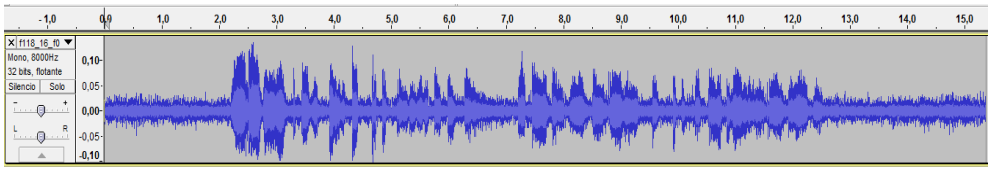


Figura 5.1. b) Fichero de voz limpia con ruido de fondo estacionario (SNR= 0 dB), MOBIO.

La pronunciación de la Figura 5.1. b) es la misma que la pronunciación de la Figura 5.1. a) pero con niveles de ruido estacionario más elevados. Un VAD de energía sería sensible a estas variaciones de la señal de ruido, y poco robusto. La energía logarítmica de una sub-banda busca solucionar este problema al ser invariante ante las variaciones de la relación señal a ruido.

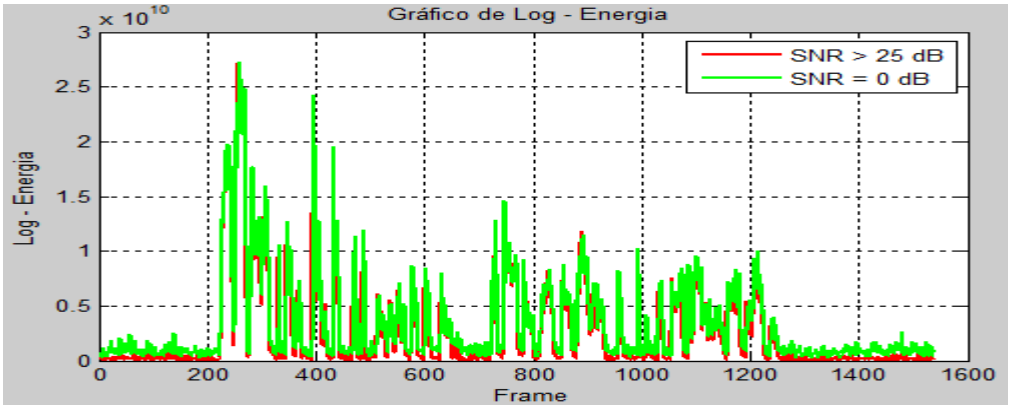


Figura 5.2. Logaritmo de la energía para diferentes SNRs, MOBIO.

Como se puede observar en la Figura 5.2, la energía logarítmica de las sub-bandas tiene similares distribuciones de energías para relaciones señal a ruido (SNR) diferentes.

A continuación, en la Figura 5.3 se observa imágenes de la respuesta del VAD para distintas SNR, de un fichero de la base MOBIO.

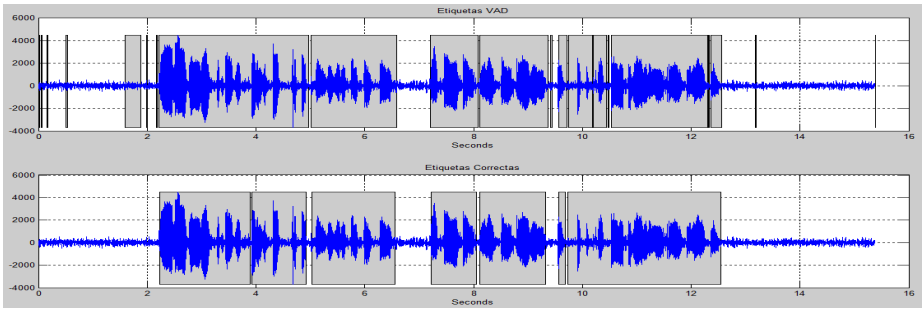


Figura 5.3. a) Respuesta del VAD para SNR de un fichero de voz nítida (SNR > 25 dB), MOBIO.

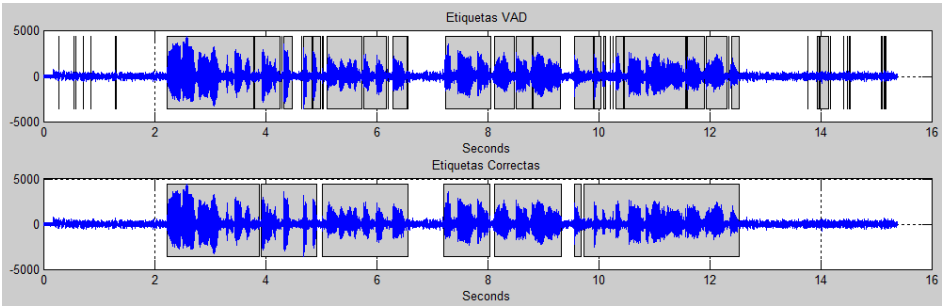


Figura 5.3. b) Respuesta del VAD para un fichero de voz con ruido de fondo (SNR= 10 dB), MOBIO.

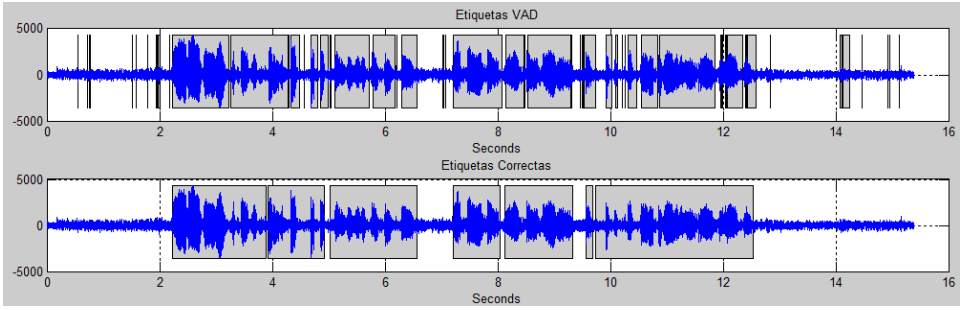


Figura 5.3. c) Respuesta del VAD para un fichero de voz con ruido de fondo (SNR= 5 dB), MOBIO.

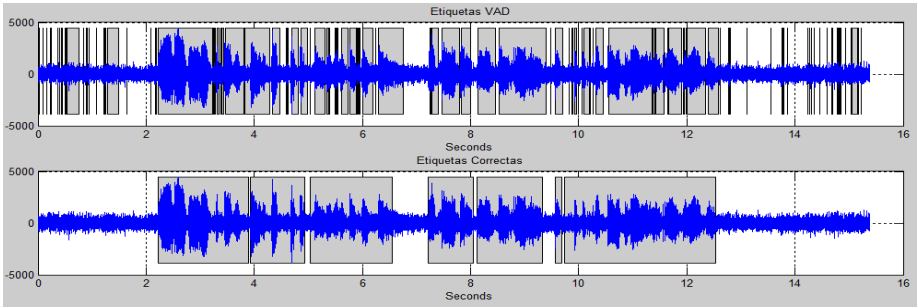


Figura 5.3. d) Respuesta del VAD para un fichero de voz con ruido de fondo (SNR= 0 dB), MOBIO.

5.2. Pruebas Iniciales de ajuste del Detector de Actividad de Voz

Para optimizar la función de respuesta del VAD, se realizan simulación donde el objetivo es lograr el máximo rendimiento variando los parámetros y las restricciones: restricciones de medias, Sintonizadores de umbral, números de canales Mel, etc. Se evalúa el desempeño del VAD utilizando la base MOBIO.

5.2.1. Coeficiente Gamma de ajuste del umbral γ

La selección adecuada de γ que aparece en la Figura 5.4; muestra la influencia del coeficiente γ, donde las curvas se obtiene de la señal a 0-dB.

Tabla 5.1. Métricas del VAD al variar el sintonizador del umbral Gamma, con ruido de conversaciones de fondo (SNR= 0 dB), MOBIO.

Escenarios	Gamma γ	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB	1	59.5	97.7	74.8
	0.95	61.4	97.2	75.8
	0.9	62.2	96.8	76.0
	0.85	63.1	96.4	76.4
	0.8	64.6	95.8	77.1
	0.75	66.8	93.2	77.4
	0.7	68.7	92.2	78.1
	0.65	70.4	89.6	78.1
	0.6	73.7	87.2	79.1
	0.55	76.1	83.6	79.1
	0.5	78.2	80.2	79.6
	0.45	81.7	77.8	79.7
	0.4	85.8	70.1	79.5
	0.35	89.4	64.4	79.4
	0.3	91.0	60.7	78.9
	0.25	93.1	53.6	77.3
	0.2	95.0	42.0	73.8
	0.15	96.6	34.1	71.6

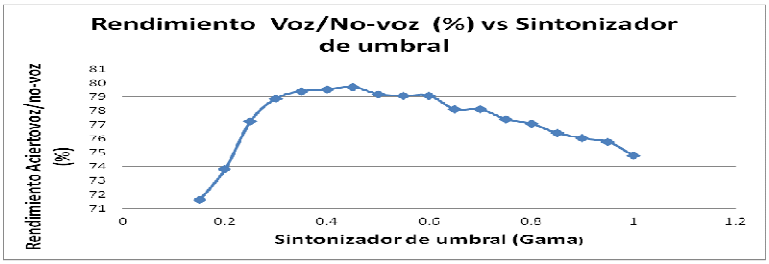


Figura 5.4. Rendimiento del VAD al variar el sintonizador del umbral Gamma, con ruido de conversaciones de fondo (SNR= 0 dB), MOBIO.

Como se puede apreciar, γ se determina como una solución de compromiso del rendimiento del VAD. A partir de este experimento, nos damos cuenta que γ= 0.45 logra un equilibrio óptimo, donde el rendimiento es alrededor de 80 %, mientras que el rendimiento disminuye a la derecha e izquierda de este valor.

5.2.2. Restricciones de medias (delta) δ

Tabla5.2. Métricas del VAD al variar la restricciones de medias, con ruido de conversaciones de fondo (SNR= 0 dB), MOBIO.

Escenarios	Delta δ	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB	1.5	80.0	76.6	78.6
	2	80.0	76.6	78.6
	2.5	80.0	76.3	78.5
	3	80.3	75.2	78.2
	3.5	81.7	74.8	78.9
	4	81.3	74.4	78.5
	4.5	81.4	73.4	78.2
	5	82.2	72.9	78.5
	5.5	79.3	73.9	77.1
	6	76.1	74.4	75.4
	6.5	74.5	74.5	74.5
	7	73.9	75.8	74.7
	7.5	73.8	75.0	74.3
	8	72.7	75.5	73.8
	8.5	71.6	75.5	73.2
	9	68.6	76.9	71.9
	9.5	64.7	78.2	70.1
	10	62.2	81.0	69.7

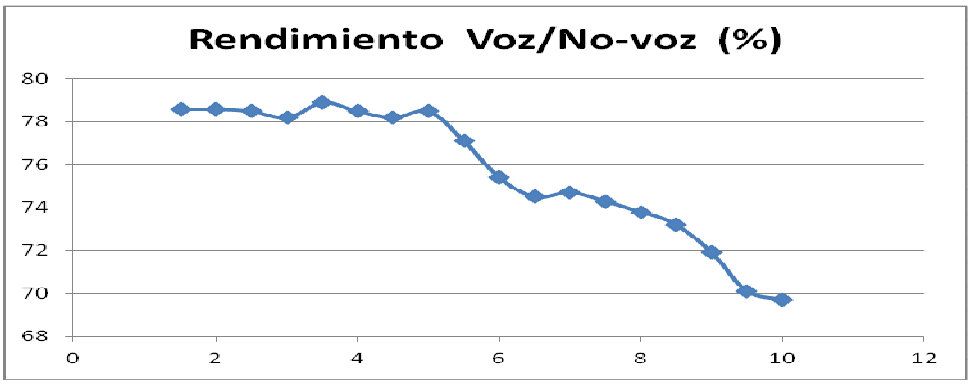


Figura 5.5. Rendimiento del VAD al variar la restricción de medias, delta, con ruido de conversaciones de fondo (SNR= 0 dB), MOBIO.

5.2.3 Rendimiento de las Sub-bandas de frecuencia Mel

Tabla 5.3.Métricas del VAD en cada sub-bandas, SNR= 0 dB, delta=0.45, 0.10, 1.00, MOBIO.

Escenarios	Sub-banda	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB delta=0.45	1	90.0	79.9	86.0
	2	86.1	85.4	85.8
	3	73.6	64.6	70.0
	4	31.3	76.5	49.4
	5	14.1	90.3	44.6
	6	25.5	87.5	50.3
	7	53.6	69.3	59.9
	8	32.6	79.9	51.5

Escenarios	Sub-banda	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB delta=0.10	1	96.0	52.4	78.6
	2	93.8	56.3	78.8
	3	85.3	48.5	70.6
	4	64.7	47.2	57.7
	5	54.6	50.8	53.1
	6	64.1	53.9	60.0
	7	75.6	49.0	65.0
	8	61.4	51.3	57.4

Escenarios	Sub-banda	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB delta=1.00	1	83.0	94.0	87.4
	2	80.4	96.6	86.9
	3	56.0	76.3	64.1
	4	13.4	97.7	47.2
	5	6.6	99.5	43.8
	6	12.0	95.9	45.6
	7	30.7	89.0	54.0
	8	13.8	96.6	46.9

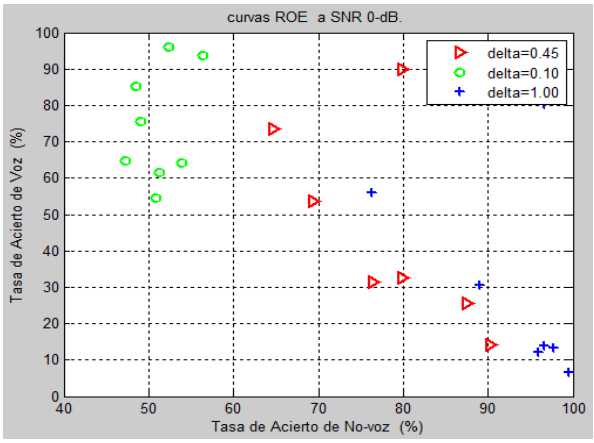


Figura 5.6. Curva ROC para diversas γ con SNR 0-dB, MOBIO.

La Figura 5.6 muestra la influencia del coeficiente γ en cada sub-bandas, donde la curvas ROC se obtienen de SNR 0 dB. A partir de este experimento, nos damos cuenta que γ de 0.45 logra un equilibrio entre la tasa de acierto de voz y tasas de acierto de no voz. Además los máximos rendimientos se observan en las primeras sub-bandas.

5.2.4. Rendimiento del VAD al variar el número de sub-banda N

Tabla 5.4.Métricas del VAD en cada de las sub-bandas SNR= 0 dB delta=0.45 N=8, MOBIO.

Escenarios	Sub-banda	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB delta=0.45 N=8	1	90.0	79.9	86.0
	2	86.1	85.4	85.8
	3	73.6	64.6	70.0
	4	31.3	76.5	49.4
	5	14.1	90.3	44.6
	6	25.5	87.5	50.3
	7	53.6	69.3	59.9
	8	32.6	79.9	51.5

Tabla 5.5. Métricas del VAD en cada de las sub-bandas SNR= 0 dB delta=0.45 N=4, MOBIO.

Escenarios	Sub-banda	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB delta=0.45 N=4	1	85.8	87.7	86.5
	2	62.1	78.7	68.7
	3	15.2	92.2	46.0
	4	43.2	82.1	58.8

Tabla 5.6. Métricas del VAD en cada de las sub-bandas SNR= 0 dB delta=0.45 N=12, MOBIO.

Escenarios	Sub-banda	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
SNR= 0 dB delta=0.45 N=12	1	90.0	56.3	76.5
	2	90.1	90.6	90.3
	3	85.2	83.1	84.3
	4	75.7	60.9	69.8
	5	40.2	65.9	50.5
	6	20.4	80.0	44.2
	7	14.7	83.0	42.0
	8	15.1	91.1	45.5
	9	47.6	73.9	58.1
	10	58.2	65.1	60.9
	11	38.1	80.0	54.9
	12	31.6	74.4	48.7

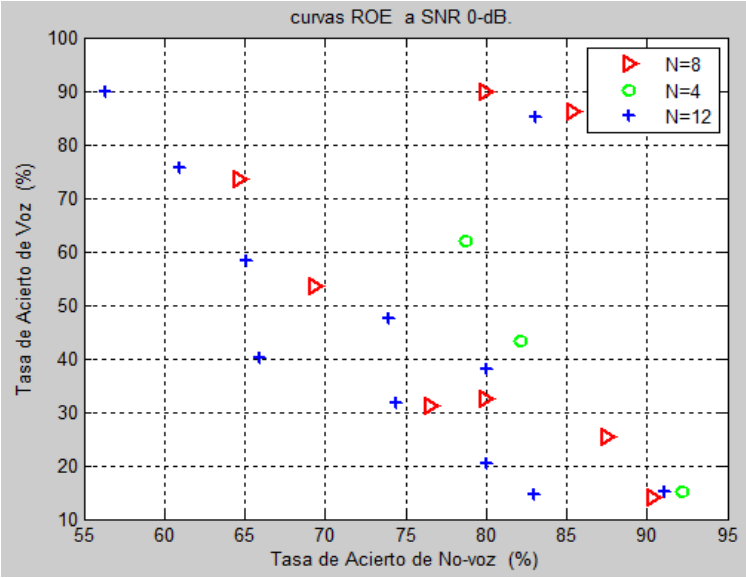


Figura 5.7. Curva ROE para varios número de sub-bandas (N) , MOBIO.

N varía de 4 a 12, mientras que otros parámetros se mantienen constantes. Al aumentar el número de sub-bandas mejora el rendimiento del VAD, porque aumenta la resolución de frecuencia. Cuándo N > 8, el rendimiento VAD mejora, pero para N > 12, no hay mejoras significativas; por lo que que N = 8, produce el mejor resultados entre costo computacional y rendimiento.

Podemos visualizar que de todos los parámetros, el coeficiente γ de ajuste del umbral de la sub-banda es el más sensible al desempeño.

El incremento del número de sub-bandas **N** influye en el rendimiento del VAD y su carga computacional aumenta.

La longitud de las ventanas de inicialización M tiene un efecto sobre la capacidad de ejecución en tiempo real; pocas muestras M benefician la capacidad computacional. Para expresiones que inician con voz, un pequeño M puede ocurrir que la señal de ruido no esté disponible para la inicialización del modelo y la distribución de la señal de ruido se inicializa de forma incorrecta por las muestras de voz. Por esta razón, **M** debe ser lo suficientemente grande para garantizar algunas muestras no-voz que se utilizarán para inicialización. Una selección adecuada sería M= 60 - 400 para ser utilizada en aplicaciones prácticas.

En la Figura 5.8, se puede observar cómo las medias de voz, no-voz, el umbral, y la probabilidad de presencia de voz (SPP) varían con las tramas que llegan al VAD.

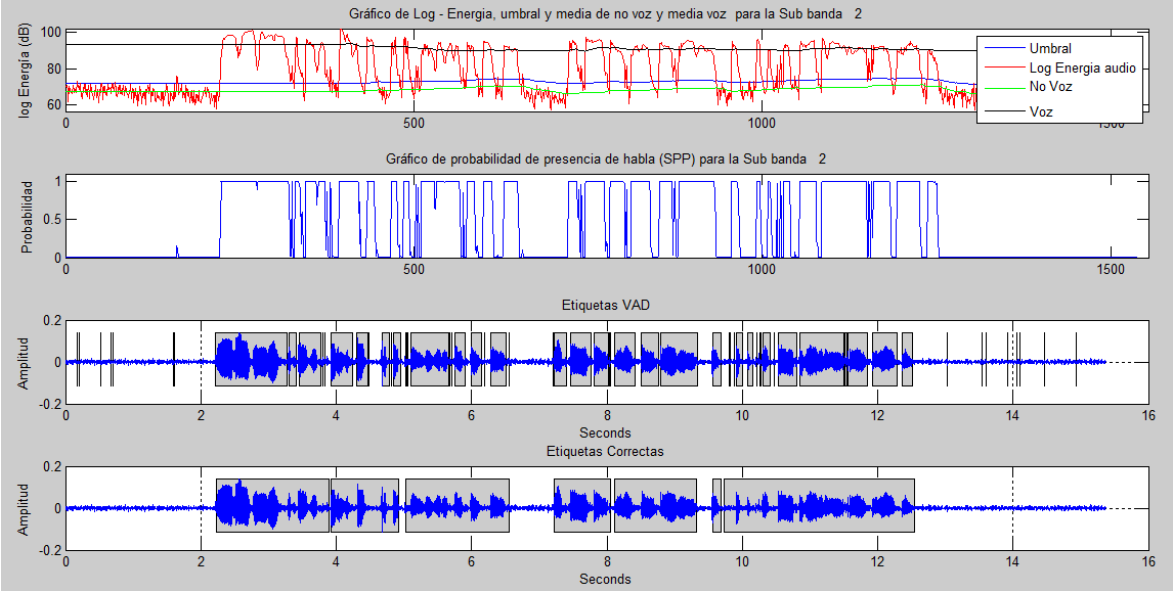


Figura 5.8. a)Energía logarítmica, umbral, media de no voz y media voz, Sub-banda 2.
b) Probabilidad de presencia de voz . c) Etiquetas VAD y d) Etiquetas Correctas, MOBIO.

5.3. Tasa de acierto de voz y tasa de acierto de no-voz del VAD

Se aplica al Detector de Actividad de Voz las base de datos de audio de 3 escenarios de ruido con 4 diferentes SNR (0dB, 5dB, 10dB, Clean); comparándose su respuesta con las etiquetas reales de las declaraciones editadas en Audacity [30] para MOBIO y TIMIT. Las marcas de inicio/fin y las señales de pausa se consideran como no-voz.

Un programa realizado en MatLab permite obtener las tasas de acierto de voz y no voz. Podemos calcular las medidas de evaluación y realizar gráficas con estas variables.

Se evalúa el desempeño del VAD utilizando la base MOBIO.

Tabla 5.7. Tasas de acierto voz/no-voz del VAD, base MOBIO.

SNR (dB)	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
Clean	97.61	91.86	95.31
10	90.79	90.90	90.83
5	86.34	88.63	87.26
0	81.69	74.83	78.94

Se puede observar que la tasa de aciertos para el VAD disminuye a medida que la SNR baja. Para la señal de audio sin añadir ruido el rendimiento del VAD es del 95.31%, la tasa de aciertos voz del 97.61% y tasa de acierto de no-voz del 91.86%. Para la señal de audio con SNR de 0 dB el rendimiento del VAD es del 78.94%, la tasa de aciertos voz del 81.69% y tasa de acierto de no-voz del 74.83%.

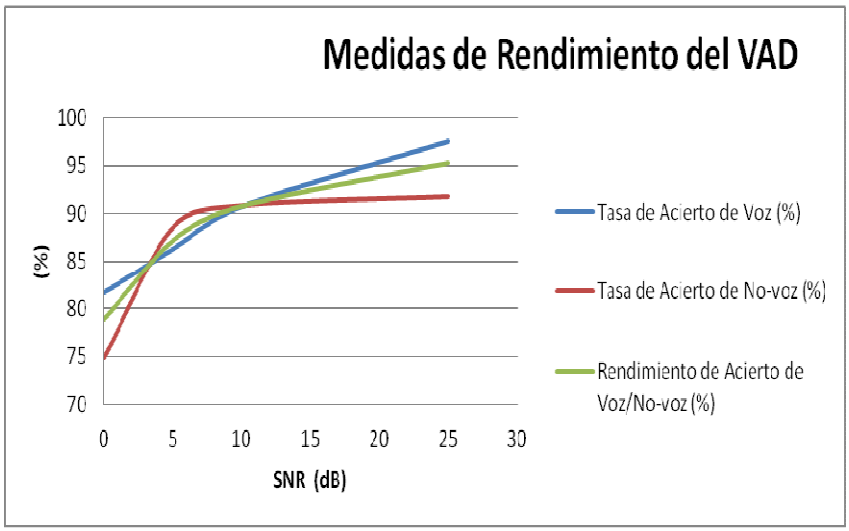


Figura 5.9. Rendimiento de tasa de acierto de voz/no-voz, base MOBIO.

Se realiza las siguientes observaciones:

El rendimiento decae un 16.37% cuando SNR disminuye desde un audio sin añadir ruido a un audio con SNR de 0 dB.

El VAD tiene un rendimiento satisfactorio en condiciones de alta SNR, ubicándose el máximo rendimiento en 95.31%. El rendimiento del VAD baja cuando disminuye la SNR; para 0 dB el rendimiento es 78.94%.

5.4. Tasas de valores perdidos de voz y Tasas de falsas alarmas del VAD

Tabla 5.8. Tasas de valores perdidos y falsas alarmas del VAD, base MOBIO.

SNR (dB)	Miss_rate (%)	False_Alarm_rate (%)
Clean	2.38	8.13
10	9.20	9.09
5	13.65	11.36
0	18.31	25.16

La tasa de falsas alarmas para la señal de audio sin añadir ruido es 8.13, mientras que la de falsos rechazos de 2.38%. Para una señal de audio con SNR de 0 dB, la tasa de falsas alarmas es 25.16%, y la de falsos rechazos de 18.31%; teniendo en cuanta los valores anteriores se obtiene una tasa de error total para SNR de 0 dB de 21.74%; a medida que la SNR disminuye, la tasa de falsa alarma y la tasa de valores perdidos aumenta.

El VAD da resultados satisfactorios en la detección de voz y en la detección de no voz, presentando una degradación leve en condiciones desfavorables de ruido, volviéndose útil en aplicaciones de procesamiento de voz.

5.5. Experimentos de Inicio de habla /no habla

Diseñamos dos experimentos para evaluar la capacidad de discriminación del VAD SGMM. El primer lugar en condiciones generales, donde el conjunto de datos satisface el supuesto de principio de "no habla al inicio de las tramas".

Se evalúa el desempeño del VAD utilizando la base TIMIT.

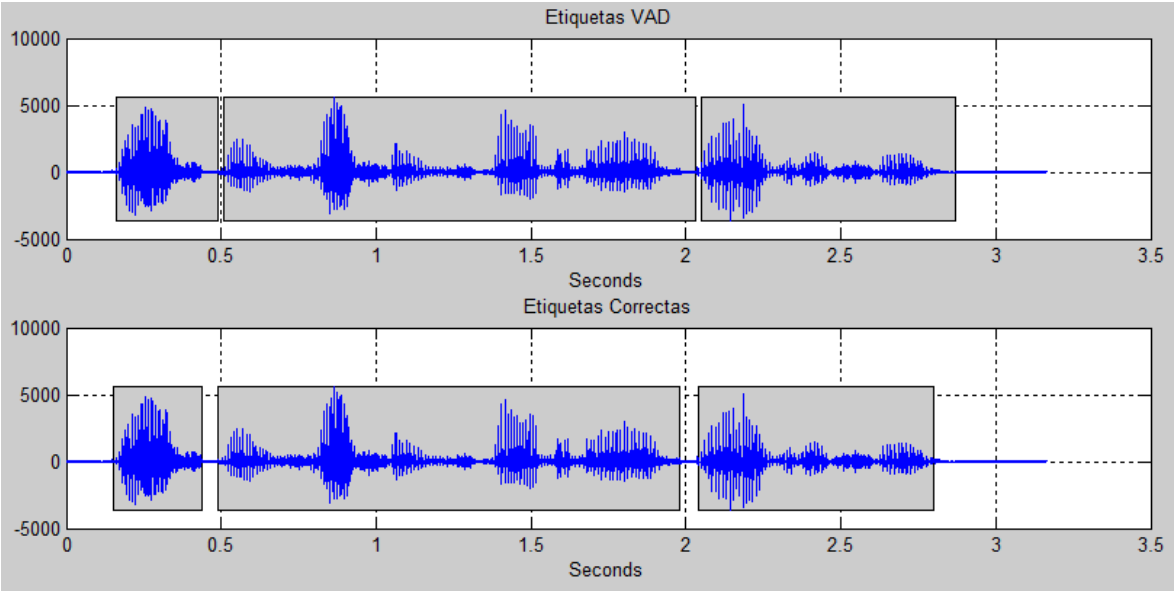


Figura 5.10. Etiquetas VAD y Etiquetas Correctas, inicio con tramas no-voz y SNR alta, TIMIT..

A continuación se muestra el estudio del comportamiento del VAD con las clases “voz” y “no voz” para audios de la base de datos TIMIT con voz sin ruido y varias SNR.

Tabla 5.9. Métricas del VAD, base TIMIT, SNR> 25 dB delta=0.45 , base TIMIT.

Escenarios	**Audio**	**Tasa de Acierto de Voz (%)**	**Tasa de Acierto de No-voz (%)**	**Rendimiento Acierto de Voz/No-voz (%)**
SNR>20 dB	test_dr1_mwbt0_sx113	89.4	60.0	85.0
	test_dr1_mwbt0_sx203	84.0	60.8	80.2
	test_dr1_mwbt0_sx23	78.2	90.0	80.4
	test_dr1_mwbt0_sx293	82.5	92.1	83.5
	test_dr1_mwbt0_sx383	74.8	93.4	77.7
	test_dr3_mlnt0_sx282	74.9	52.0	73.1
	test_dr4_mlll0_sx193	84.9	34.4	61.3
	test_dr4_mlll0_sx283	95.5	61.9	86.9
	test_dr4_mtls0_sx290	62.1	94.4	63.9
	test_dr5_mbpm0_sx137	83.8	14.9	65.1
	test_dr5_mklt0_sx313	76.9	88.1	80.4
	test_dr6_mcmj0_sx284	59.1	71.9	61.1
	test_dr6_mjdh0_sx274	80.5	46.8	70.4
	test_dr6_mjdh0_sx364	76.0	100.0	77.4
	test_dr7_mgrt0_sx100	62.6	51.7	58.1
	test_dr7_mgrt0_sx370	96.6	31.8	88.1
	test_dr7_mnjm0_sx230	73.4	75.8	73.6
	test_dr7_mnjm0_sx320	86.4	92.2	87.1
	test_dr7_mnjm0_sx50	62.5	100.0	66.5
	test_dr8_mjln0_sx279	74.4	57.4	72.2
	test_dr8_mjln0_sx99	78.4	66.7	76.6
	test_dr8_mpam0_sx199	82.0	100.0	83.5

Tabla 5.10. Métricas Promedio del VAD, base TIMIT, SNR> 25 dB delta=0.45

Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
96.8	63.6	90.3

Se puede observar que la tasa de aciertos para el VAD, utilizando la base TIMIT y la señal de audio sin añadir ruido, el rendimiento promedio es 90.3%, la tasa de aciertos voz del 96.8% y la tasa de acierto de no-voz del 63.6%.

Se repite los experimentos con varias SNR, obteniendo los siguientes valores:

Tabla 5.11. Métricas Promedio del VAD, base TIMIT, para diferentes SNR.

SNR (dB)	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)	Miss_rate (%)	False_Alarm_rate (%)
Clean	96.8	63.6	90.3	3.2	36.4
10	77.1	91.3	78.8	22.9	8.7
5	74.7	88.9	76.3	25.3	11.1
0	78.1	69.8	75.1	21.9	30.2

Se realiza las siguientes observaciones:

Para la señal de audio con SNR de 0 dB el rendimiento del VAD es del 75.1%, la tasa de aciertos voz del 78.1% y tasa de acierto de no-voz del 69.8%.

El VAD tiene un rendimiento satisfactorio en condiciones de alta SNR, ubicándose el máximo rendimiento en 90.3%. El rendimiento del VAD baja cuando disminuye la SNR; para 0 dB el rendimiento es 75.1%.

El rendimiento decae un 15.2% cuando SNR disminuye desde un audio sin añadir ruido a un audio con SNR de 0 dB.

La tasa de falsas alarmas para la señal de audio sin añadir ruido es 36.4%, mientras que la de falsos rechazos de 2.38%. Para una señal de audio con SNR de 0 dB, la tasa de falsas alarmas es 30.2%, mientras que la de falsos rechazos de 21.9%; teniendo en cuanta los valores anteriores se obtiene una tasa de error total para SNR de 0dB de 26.05%.

5.5.1. Curvas ROE para la base de datos TIMIT.

Tabla 5.12. Tasa de Valores Perdidos de Voz y Tasa de Aciertos de No-voz, TIMIT

SNR (dB)	Miss_rate (%)	Tasa de Acierto de No-voz (%)
Clean	3.2	63.6
10	22.9	91.3
5	25.3	88.9
0	21.9	69.8

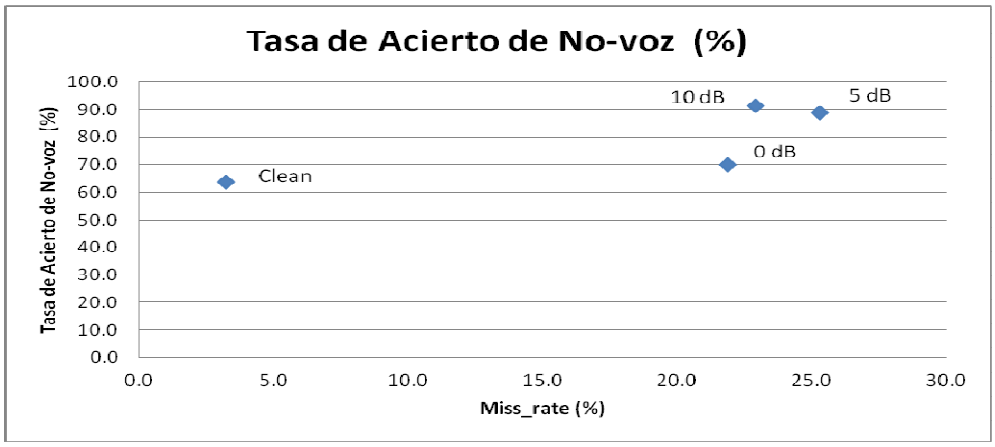


Figura 5.11. Tasa de Acierto de No-voz vs Miss rate , base TIMIT

En la Figura 5.11, se ilustran los resultados del rendimiento del VAD, para la base TIMIT y el efecto que tiene SNR en la detección. Se puede deducir que cuando la señal no está afectada con ruido, la tasa de valores perdidos es baja (3.2%) y, a medida que disminuye la SNR el punto de operación se desplaza hacia la derecha, aumentando la tasa de valores perdidos (21.9 para 0 dB).

Tabla 5.13. Tasa de Aciertos de Voz y Tasa de Aciertos de No-Voz, TIMIT

SNR (dB)	Tasa de Acierto de No-voz (%)	Tasa de Acierto de Voz (%)
Clean	63.6	96.8
10	91.3	77.1
5	88.9	74.7
0	69.8	78.1

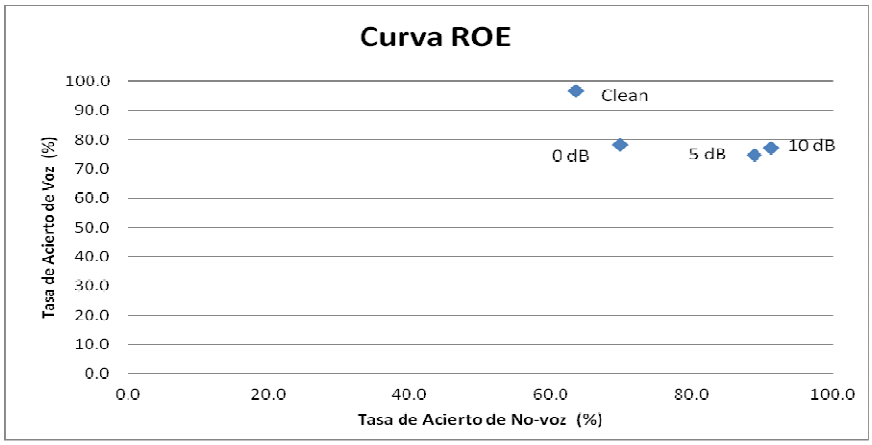


Figura 5.12. Tasa de Acierto de Voz vs Tasa de Acierto de No-voz, base TIMIT

La Figura 5.12. se observa el efecto que tiene SNR en la detección. Se puede observar que cuando la señal de audio no está afectada con ruido, su tasa de aciertos de voz es alta (96.8%), a medida que disminuye la SNR el punto de operación se desplaza hacia la abajo, lo que disminuye la tasa de aciertos de voz.

5.5.2 Experimentos con voz al inicio de las tramas

El propósito del segundo experimento es evaluar la influencia del "principio de no habla" sobre el rendimiento del VAD, cortando las tramas iniciales de expresiones de silencio de señal. Por lo tanto, algunas expresiones empezarán con señal de voz. El resultado experimental se muestra en la Figura 5.13.

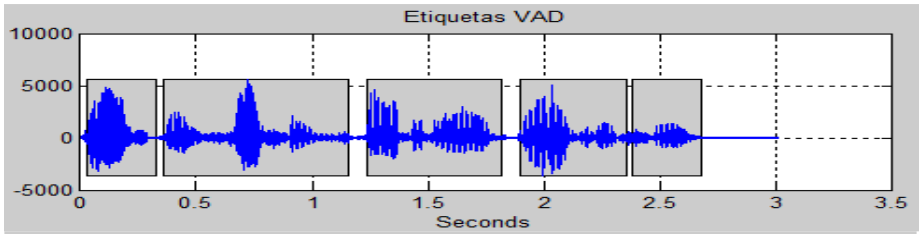
El archivo de audio que se utiliza para la prueba es: train_dr1_mwbt0_sx23.wav

Los parámetros utilizados en la implementación del algoritmo, con una tasa de muestreo de 8 KHz se encuentra en la Tabla 5.14.

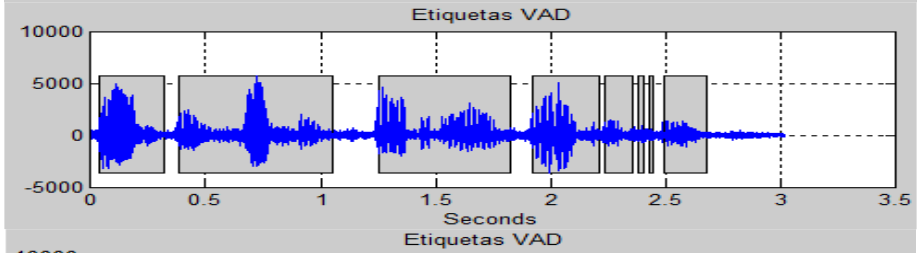
Tabla 5.14. Valores de los parámetros utilizados en la implementación del VAD_SGMM, inicio con voz.

Longitud de las ventanas de inicialización	400
Tamaño de cada ventana o frame	20 ms
Solapamiento	50%
Frecuencia de muestreo	8000 Hz
Longitud de la ventana FFT	256
Números de canales Mel	8
Calibrador del umbral Gamma	0.45
Umbral de Votación	2 (2 de 8)
Parámetros hangover:	bth=4 hcnt=5
Coeficientes de pesos de actualización (Alfa)	0.99
Restricciones de medias (delta)	3.5

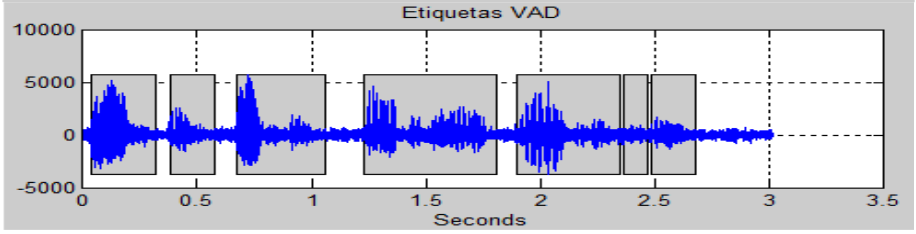
Clean



10 dB



5 dB



0 dB

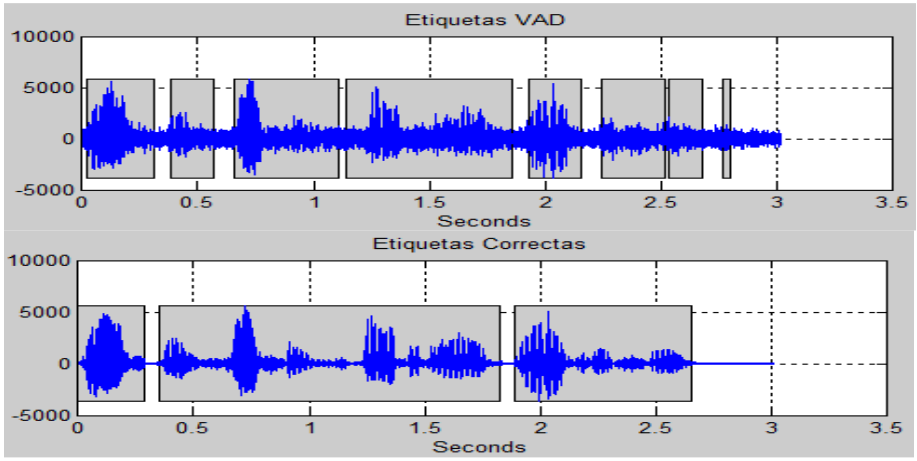


Figura 5.13 Respuesta del VAD para varias SNR y Etiquetas Correctas, inicio con voz, base TIMIT.

La Figura 5.13 muestra el comportamiento del VAD si la señal de audio empieza con voz para varias SNR, observando que cuando la señal de audio es limpia, su respuesta de rendimiento es apropiada; a medida que la SNR decrece, se aprecia ligeros errores en su respuesta, validando la robustez del VAD para habla al principio de las tramas.

Tabla 5.15. Tasa de aciertos de voz, tasa de aciertos de no de voz y rendimiento del VAD para diferentes SNR, inicio con voz, base TIMIT.

SNR (dB)	Tasa de Acierto de Voz (%)	Tasa de Acierto de No-voz (%)	Rendimiento Acierto de Voz/No-voz (%)
Clean	94.92	87.2	93.72
10	85.93	89.36	86.46
5	86.32	89.36	86.79
0	89.45	76.59	87.45

En este experimentos con señal de audio limpia que inicia con voz, el rendimiento del VAD es del 93.72%, la tasa de aciertos voz del 94.92% y tasa de acierto de no-voz del 87.23%.

Para la señal de audio con SNR de 0 dB el rendimiento del VAD es del 87.45%, la tasa de aciertos voz del 89.45% y tasa de acierto de no-voz del 76.59%.

Como se puede ver, el rendimiento decae un 4.27% cuando SNR disminuye desde un audio sin añadir ruido a un audio con SNR de 0 dB; las métricas validan la robustez del VAD para habla al principio de las tramas.

En la señale sin añadir ruido la tasa de acierto de no voz, es menor que las tasas de acierto de voz, esto se debe a que en la calibración del VAD_SGMM, el parámetro del sintonizador de umbral está configurado en γ = 0.45, cuyo objetivo es evitar perder los segmentos de voz, en deterioro de la efectividad en la detección de los segmentos de silencios.

Los experimentos descritos aportan medidas del rendimiento del VAD, suministrando información relevante acerca del rendimiento del VAD y puede usarse para optimizar su operación.

5.6 Comparativas del VAD_SGMM con el VAD basado en Energía

En este trabajo, se han comparado el VAD_SGMM con el VAD basado en Energía utilizando la base de datos MOBIO, en varios escenarios de SNR (clean, 10dB, 5dB y 0dB).

Tabla 5.16. Rendimiento acierto de voz/no-voz VAD basado en Energía y VAD_SGMM, base MOBIO.

SNR (dB)	Rendimiento (%) VAD_SoX	Rendimiento (%) VAD_1	Rendimiento (%) VAD_Energía	Rendimiento (%) VAD_SGMM
Clean	81.4	79.1	81.83	95.31
10	78.2	82.9	79.29	90.83
5	75.7	79.7	77.36	87.26
0	71.1	77.9	73.59	78.94

Sobre la base de la prueba establecida, el VAD_SGMM tiene rendimientos más altos para ruidos y SNR utilizadas.

Así, para una señal de audio sin añadir ruido el VAD_SGMM tiene un rendimiento de 95.31 mientras que para el VAD_Energía, el rendimiento es de 81.83%. Para una señal de audio con SNR de 0 dB el VAD_SGMM tiene un rendimiento de 78.94% y el VAD_Energía un rendimiento de 73.59%.

Por lo tanto, el VAD_SGMM tiene en promedio rendimientos superiores a un 10% que el VAD_ Energía.

Tabla 5.17. Tasa de Acierto de Voz VAD_Energía y VAD_SGMM, base MOBIO.

SNR (dB)	Tasa de Acierto de Voz (%) VAD_SoX	Tasa de Acierto de Voz (%) VAD_1	Tasa de Acierto de Voz (%) VAD_Energía	Tasa de Acierto de Voz (%) VAD_SGMM
Clean	60.9	97.1	60.78	97.61
10	56.4	93.1	55.67	90.79
5	53.6	91.5	52.8	86.34
0	49.4	88.1	48.02	81.69

Para la señal de audio sin añadir ruido el VAD_SGMM tiene una tasa de acierto de voz de 97.61%; para el VAD_Energía, la tasa de acierto de voz es de 60.78%. Si la señal de audio tiene una SNR de 0 dB el VAD_SGMM tiene la tasa de acierto de voz de 81.69, mientras la del VAD_Energía es de 48.02%.

Por lo tanto, el VAD_SGMM tiene en promedio una tasa de acierto de voz superiores a un 35% que el VAD_Energía.

Tabla 5.18. Tasas de valores perdidos y falsas alarmas VAD_Energía y VAD_SGMM, base MOBIO.

SNR (dB)	Miss_R (%) VAD_SoX	Miss_R (%) VAD_1	Miss_R (%) VAD_Energía	Miss_R (%) VAD_SGMM	False_Alarm_R (%) VAD_SoX	False_Alarm_R (%) VAD_1	False_Alarm_R (%) VAD_Energía	False_Alarm_R (%) VAD_SGMM
Clean	39.1	2.9	17.9	2.4	2.4	33.0	8.8	8.1
10	43.6	6.9	44.3	9.2	4.5	23.7	2.0	9.1
5	46.4	8.5	47.2	13.6	6.5	27.5	2.9	11.7
0	50.6	11.9	52.0	18.3	11.1	29.2	5.6	25.2

Los resultados de los ensayos muestra que el VAD_SGMM presenta tasas de valores perdidos más bajas comparado con el VAD basado en energía.

Para una señal de audio sin añadir ruido el VAD_SGMM tiene una tasa de valores perdido de 2.4% y el VAD_Energía de 17.9%; la tasa de falsas alarmas del VAD_SGMM es de 8.1%, mientras la del VAD basado en energía es de 8.8%.

Para una señal de audio con SNR de 0 dB el VAD_SGMM tiene una tasa de valores perdidos de 18.3% y el VAD_Energía de 52.0%; la tasa de falsas alarmas del VAD_SGMM es de 25.2%, mientras que el VAD_Energía es de 5.6%.

Consecuentemente, el VAD_SGMM tiene en promedio tasas de valores perdidos menores a un 30% que el VAD basado en Energía.

Capítulo 6

Conclusiones y Trabajos Futuros

A continuación se presentan las conclusiones y líneas futuras de investigación.

6.1. Conclusiones

1. En este trabajo, se ha analizado e implementado un Detector de Actividad de Voz no Supervisado, robusto a entornos ruidosos y se presenta los resultados obtenidos en los experimentos con las bases de datos MOBIO y TIMIT; destacándose el VAD_SGMM, por su factible implementación en sistemas de detección y también por su aplicación en tiempo real.

2. Escogiendo las tramas de inicialización, el VAD_SGMM, construye los modelos de voz y no-voz correctamente, verificando su funcionamiento sin supervisión en la detección de expresiones de audio que empiezan con tramas de voz y en expresiones de audio que inician con tramas de no-voz.

3. Los parámetros utilizados en el modelo deben ser regulados en base a pruebas de afinamiento, examinando rendimientos y tasas de errores; para nuestros experimentos los máximos rendimientos se han logrado si el VAD_SGMM opera con sintetizador de umbral γ= 0.45 y restricción de media δ = 3.5.

4. El VAD_SGMM tiene un rendimiento satisfactorio en condiciones de alta SNR, para la base TIMIT el máximo rendimiento es 90.3%. El rendimiento del VAD decrece cuando disminuye la SNR; a 0 dB el rendimiento es 75.1%, concluyendo que el rendimiento disminuye en un 15.2% cuando SNR cambia de un audio sin añadir ruido a un audio con SNR de 0 dB.

5. El VAD_SGMM con la base TIMIT tiene tasas de falsas alarmas para la señal de audio sin añadir ruido de 36.4%, mientras que la tasas de falsos rechazos es 3.2%. Para una señal de audio con SNR de 0 dB, se incrementa la tasa de falsas alarmas a 30.2% y la tasas de falsos rechazos a 21.9%; considerando los valores anteriores se obtiene una tasa de error total para SNR de 0dB de 26.05%.

6. Con la Base MOBIO, el VAD_SGMM tiene en promedio rendimientos (voz/no-voz) superiores a un 10% respecto al VAD basado en energía, de igual manera, el promedio de la tasa de acierto de voz supera en un 35%.

Las aportaciones obtenidas en este trabajo son:

- Se ha realizado el etiquetado manual de los audio de prueba, generando las marcas en el tiempo de inicio y terminación de frases de voz para 43 audio de la base MOBIO y 23 audio de prueba para la base TIMIT; a los cuales se han añadido tres tipos de ruido de conversaciones de fondo con niveles de SNR: alto (10 dB), medio (5 dB) y bajo (0 dB).

- Los procesos del VAD_SGMM se ha implementado en programación en MatLab, replicando y configurando para entornos reales ruidosos.

- Para contrastar los resultados obtenidos del VAD_SGMM, se ha realizado comparativas con el VAD basado en energía.

6.2. Trabajos Futuros

Los posibles trabajos futuros irán encaminados a perfeccionar las técnicas presentadas:

- Utilizar el VAD_SGMM en sistemas de reconocimiento de habla para su aplicación real y realizar investigaciones en sistemas robustos de reconocimiento de voz.

- Analizar otras variables de la señal de audio que presenten inmunidad al ruido, e implementar un modelo sin supervisión, mediante el uso de funciones multivariadas.

- Implementar Detectores de Actividad de Voz mediante técnicas de aprendizaje como Redes Neuronales, Modelos Ocultos de Markov o Máquinas de Vectores de Soporte.

Referencias Bibliográficas

[1] J.M. Górriz. “New Advances in Voice Activity Detection using Higher Order Statistics and Optimization Strategies”. Ph. D. European Thesis by the University of Granada, Spain (Jul. 2006).

[2] D. Ying, Y. Yan, J. Dang, F. K. Soong. Voice Activity Detection Based on an Unsupervised Learning Framework. IEEE Transactions on Audio, Speech, and Language Processing 19(8), 2011.

[3] J. Chang, N. Kim, and S. Mitra, “Voice activity detection based on multiple statistical models,” IEEE Trans. Signal Process., vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[4] G. Evangelopoulos and P. Maragos. Multiband modulation energy tracking for noisy speech detection. Audio, Speech, and Language Processing, IEEE Transactions on, 14(6):2024 _2038, November 2006.

[5] Z. Nasibov, “Decision fusion of voice activity detectors,” Master’s thesis, University of Eastern Finland, Joensuu, May 2012.

[6] Sangwan A. Chiranth M.C., Jamadagni H.S., Sah R., Prasad R.V., Gaurav V., “VAD techniques for real-time speech transmission on the Internet”, IEEE Int. Conf. on High-Speech Networks and Multimedia Comm., 2002, pp 46-50.

[7] Itoh K., Mizushima M., “Environmental noise reduction based on speech/non-speech iden-tification for hearing aids”, Int. Conf. on Acoust. Speech Signal Proc., vol. 1, 1997, pp 419-422.

[8] Sohn J., Kim N. S., and Sung W., “A statistical model-based voice activity detection”, IEEE Signal Proc. letters, Jan 1999, vol. 6, no. 1, pp 1-3.

[9] Óscar Varela Serrano, Técnicas de análisis, caracterización y detección de señales de voz en entornos acústicos adversos, Madrid, 2011

[10] L. R. Rabiner, M. R. Sambur. “An algorithm for determining the endpoints of isolated utterances”. The Bell System Technical Journal. Volumen 54, nº 2, Febrero 1975. Pp. 297-315.

[11] E. Dong, G. Liu, Y. Zhou, and X. Zhang. Applying Support Vector Machines to Voice Activity Detection. In Proceedings of the International Conference on Signal Processing (ICSP), 2002.

[12] T. Kinnunen, E. Chernenko, M. Tuononen, P. Franti, and H. Li. Voice activity detection using MFCC features and support vector machine. In Proceedings of the International Conference on Speech and Computer, 2007.

[13] O. Chapelle, B. Schölkopf, and A. Zien, Semi-Supervised Learning. Cambridge, MA: MIT Press, 2006, ch. 1, p. 1–12.

[14] J. Sohn, N. Soo, and W. Sung. A statistical model-based voice activity detection. IEEE Signal Processing Letters 6(1), 1999.

[15] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, “ITU-T recommendation G.729 annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,” Communications Magazine, IEEE, vol. 35, no. 9, pp. 64 –73, sep 1997.

[16] ITU, “Coding of speech at 8 Kbit/s using conjugate structure algebraic code-excited linear –prediction (CS-ACELP). Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70,” International Telecommunication Union, 1996.

[17] ETSI, “Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels,” ETSI EN 301 708 Recommendation, 1999.

[18] Ivan J. Tashev. Sound Capture and Processing: Practical Approaches. Wiley, 2009.

[19] José C. Segura, Javier Ramírez, Carmen Benítez, Ángel de la Torre, Antonio Rubio, “Improved feature extraction based on spectral noise reduction and nonlinear feature nonlinear feature normalization”, Eurospeech 2003, pp. 353-356.

[20] http://www.slideshare.net/sofilu55/filtros-wiener, Filtros wiener, by Sofia Asadovay, 2011

[21] Izhak Shafran & Richard Rose, “Robust Speech Detection and Segmentation for Real-Time ASR Applications”, ICASSP 2003, pp. 432-434.

[22] F. Lamel, R. Rabiner, E. Rosenberg, and G. Wilpon, “An improved endpoint detector for isolated word recognition,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-29, no. 4, pp. 777–785, Aug. 1981.

[23] Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 050 Rec., ETSI, 2002.

[24] K. Ishizuka and T. Nakatani, “Study of noise robust voice activity detection based on periodic component to aperiodic component ratio,” in Proc. SAPA’06, Pittsburgh, PA, 2006, pp. 65–70.

[25] Md Sahidullah, Student Member, IEEE, Goutam Saha, Member, IEEE , Comparison of Speech Activity Detection Techniques for Speaker Recognition, arXiv:1210.0297v2 [cs.MM] 6 Oct 2012.

[26] Jonathan Kola, Carol Espy-Wilson and Tarun Pruthi "Voice Activity Detection", MERIT BIEN 2011 Final Report, pag 1 - 6.

[27] Alan Davis, Student Member, IEEE, Sven Nordholm, Senior Member, IEEE, and Roberto Togneri, Senior Member, IEEE, " Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold ", IEEE Transactions on AUDIO, Speech and Language Processing, Vol. 14, No. 2, March 2006.

[28] Chris McCool, Sébastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matějka, Jan Černocký, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Lévy, Driss Matrouf, Jean-François Bonastre, Phil Tresadern, and Timothy Cootes, “Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data”, in IEEE ICME Workshop on Hot Topics in Mobile Mutlimedia, 2012.

[29] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren , and Victor Zue; " TIMIT Acoustic-Phonetic Continuous Speech Corpus "; Linguistic Data Consortium, Philadelphia.

[30] http://audacity.sourceforge.net/

[31] Ron Kohavi and Foster Provost. Glossary of terms. Mach. Learn., 30(2- 3), 1998.

[32] Javier González-Domínguez, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano, and Joaquín González-Rodríguez. ATVS-QUT NIST SRE 2012 System Description. ATVS Biometric Recognition Group, Universidad Autónoma de Madrid, Spain.

[33] http://sox.sourceforge.net/ - open source